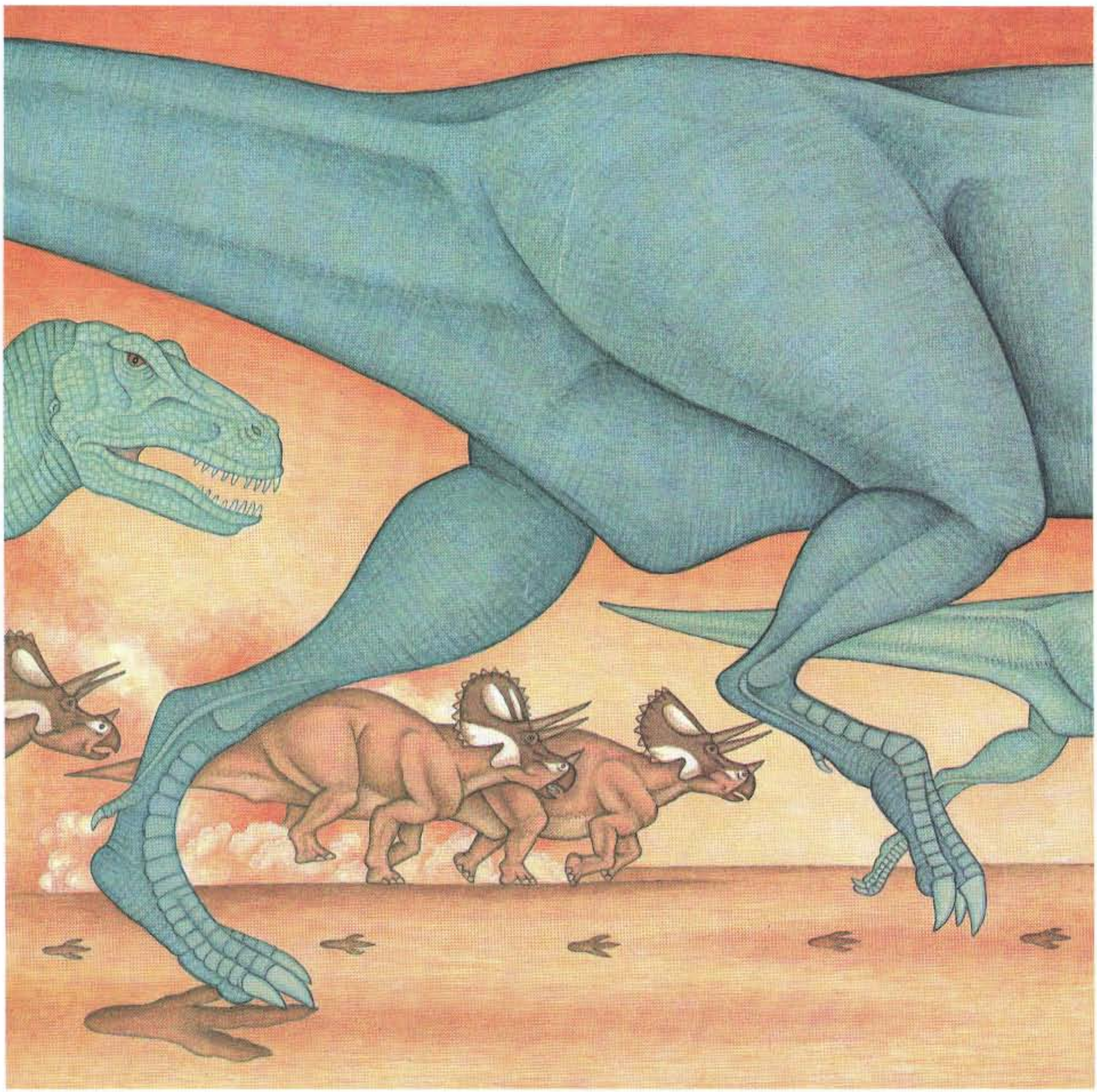
SCIENTIFIC AMERICAN

APRIL 1991
$3.50

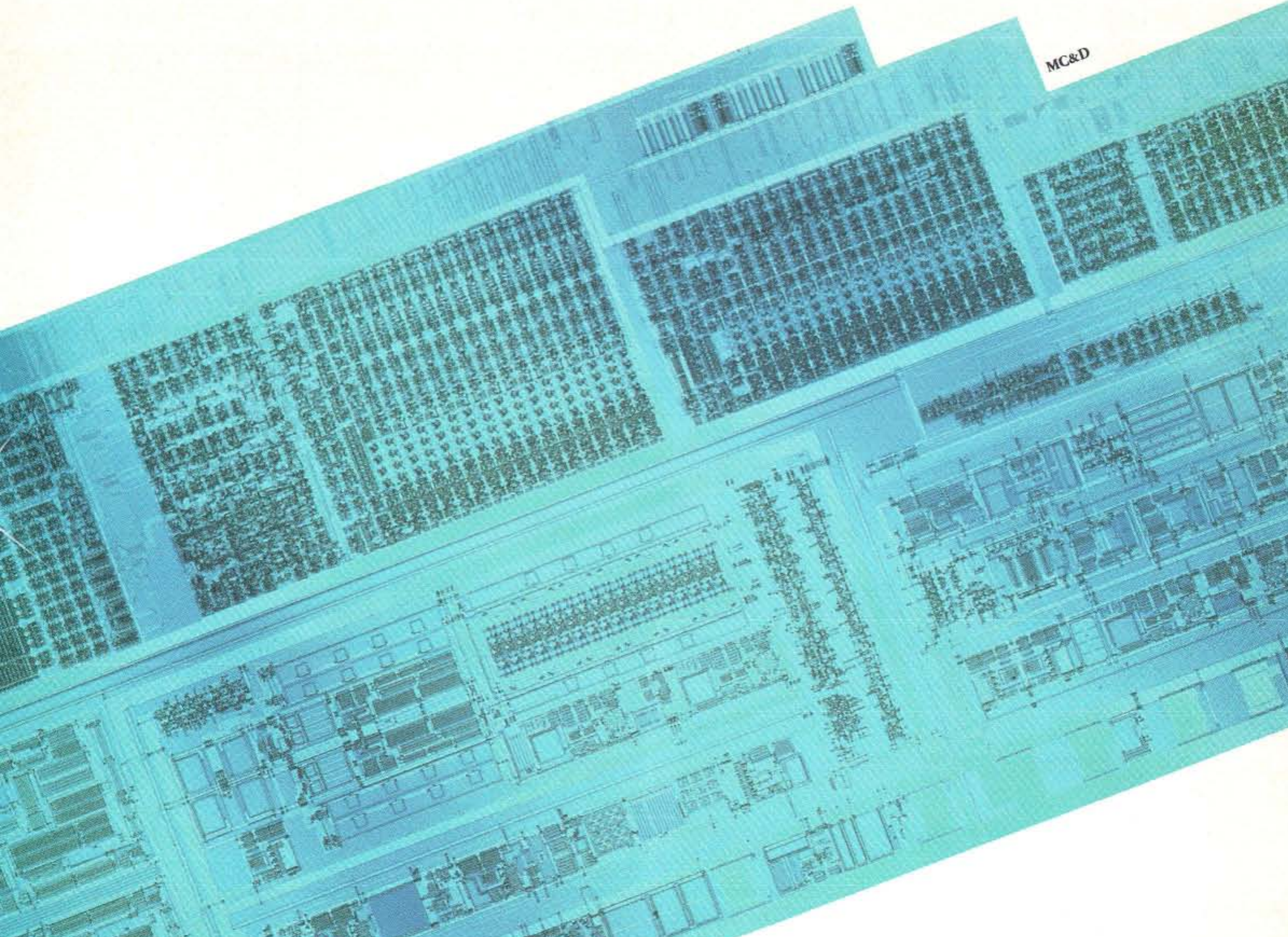
Shock waves in the void of space.

The echoes of ancient languages.

What does energy really cost?



If tyrannosaurs ran, *how fast did they go? The art of shipbuilding provides a clue.*



This microchip converts analog signals into digital signals – essential for the incorporation of ISDN into the existing telephone network. In ISDN, the new form of integrated communication, conversations, texts, images and data are sent via a single telephone cable.

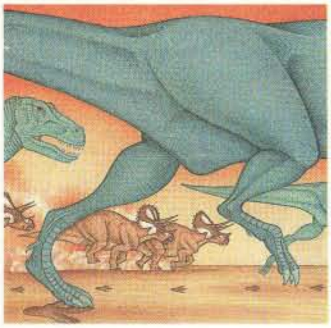
A19100-F1-Z3-X-7600

SCIENTIFIC AMERICAN

April 1991 Volume 264 Number 4

Scientific American (ISSN 0036-8733), published monthly by Scientific American, Inc., 415 Madison Avenue, New York, N.Y. 10017. Copyright © 1991 by Scientific American, Inc. All rights reserved. Printed in the U.S.A. No part of this issue may be reproduced by any mechanical, photographic or electronic process, or in the form of a phonographic recording, nor may it be stored in a retrieval system, transmitted or otherwise copied for public or private use without written permission of the publisher. Second-class postage paid at New York, N.Y., and at additional mailing offices. Authorized as second-class mail by the Post Office Department, Ottawa, Canada, and for payment of postage in cash. Canadian GST No. R 127387652. Subscription rates: one year $36 (outside U.S. and possessions add $11 per year for postage). Subscription inquiries: U.S. only 800-333-1199; other 515-247-7631. Postmaster: Send address changes to Scientific American, Box 3187, Harlan, Iowa 51537.

DEPARTMENTS



THE COVER painting depicts the carnivorous dinosaur *Tyrannosaurus* and, in the background, the herbivore *Triceratops*. How quickly these massive creatures of the Mesozoic era moved remains a mystery since they cannot be directly observed. But careful calculations, based on the principles of physics and structural engineering, could yield an answer (see "How Dinosaurs Ran," by R. McNeill Alexander, page 62).

THE ILLUSTRATIONS

Cover painting by Patricia J. Wynne

Page	Source
19	Witt/SIPA
20–21	Jason Küffer (*top*), Runk Schoenberger, Grant Heilman Photography, Inc. (*bottom*)
22	Jason Küffer
24	Kenji Hiraga, Tohoku University
25	An Pang Tsai, Akihisa Inoue and Tsuyoshi Masumoto, Tohoku University
26–31	Laurie Grace
33	Steven Lanier McKnight
34	George V. Kelvin
35	Steven Lanier McKnight (*left*), George V. Kelvin (*right*)
36	George V. Kelvin (*top and bottom*), bottom courtesy of Ted Abel and Tom Maniatis, Harvard University
37	George V. Kelvin (*top*), Jason Küffer (*bottom*)
38	George V. Kelvin
41	George Retseck, courtesy of Bruce Tsurutani, Jet Propulsion Laboratory, and Paul Rodriguez, Naval Research Laboratory
42–43	George Retseck
44	Stewart Moses, TRW (*top*), Dennis Papadopoulos and Peter Cargill, University of Maryland at College Park (*bottom*)
45	George Retseck (*top*), Charles F. Kennel (*bottom*)
46	George Retseck (*left*), Mount Wilson and Palomar Observatories (*right*)
48	Pierre Tiollais
49	Tomo Narashima
50	Jason Küffer
51–53	Tomo Narashima
56–57	Corning, Inc.
58	Michael Goodman
59	Michael Goodman (*top left*), Corning, Inc. (*top right*), Frank J. Borkowski, Foto Forum (*bottom*)
60	Corning, Inc. (*top left*), Peter Aldridge and Robin Revell, Corning, Inc. (*top right*), Edward Keating (*bottom*)
61	Corning, Inc.
62–64	Patricia J. Wynne
65	American Museum of Natural History
66–68	Patricia J. Wynne
70–71	Ian Worpole
72–73	Laurie Grace
74–75	Tom Maday (*left*), Moshe Milner/Sygma (*center*), Tom Zimberoff (*right*)
77	Jason Küffer
79	Hank Iken
89–90	Thomas C. Moore

SCIENTIFIC AMERICAN®

Established 1845

EDITOR: Jonathan Piel

BOARD OF EDITORS: Alan Hall, *Executive Editor;* Michelle Press, *Managing Editor;* Timothy M. Beardsley; Elizabeth Corcoran; Deborah Erickson; Marguerite Holloway; John Horgan; Philip Morrison, *Book Editor;* Corey S. Powell; John Rennie; Philip E. Ross; Ricki L. Rusting; Russell Ruthen; Gary Stix; Paul Wallich; Philip M. Yam

ART: Samuel L. Howard, *Art Director;* Joan Starwood, *Design Director;* Edward Bell, *Associate Art Director;* Nisa Geller, *Photos;* Johnny Johnson

COPY: Maria-Christina Keller, *Copy Chief;* Nancy L. Freireich; Jonathan Goodman; Daniel C. Schlenoff

PRODUCTION: Richard Sasso, *Vice President Production and Distribution; Managers:* Carol Albert, *Prepress;* Tanya DeSilva, *Projects;* Carol Eisler, *Manufacturing and Distribution;* Carol Hansen, *Composition;* Madelyn Keyes, *Systems;* Leo J. Petruzzi, *Manufacturing and Makeup;* Carl Cherebin; William Sherman

CIRCULATION: Lorraine Leib Terlecki, *Circulation Director;* Cary Zel, *Circulation Manager;* Rosa Davis, *Fulfillment Manager;* Katherine Robold, *Assistant Business Manager*

ADVERTISING: Robert F. Gregory, *Advertising Director.* OFFICES: NEW YORK: William Buchanan; Peter Fisch; Michelle Larsen; Meryle Lowenthal. CHICAGO: 333 N. Michigan Avenue, Chicago, IL 60601; Patrick Bachler, *Advertising Manager.* DETROIT: 3000 Town Center, Suite 1435, Southfield, MI 48075; Edward A. Bartley, *Detroit Manager;* William F. Moore. WEST COAST: 1650 Veteran Avenue, Suite 101, Los Angeles, CA 90024; Kate Dobson, *Advertising Manager;* Joan Berend, San Francisco. CANADA: Fenn Company, Inc. DALLAS: Griffith Group.

ADVERTISING SERVICES: Laura Salant, *Sales Services Director;* Diane Schube, *Promotion Manager;* Mary Sadlier, *Research Manager;* Ethel D. Little, *Advertising Coordinator*

INTERNATIONAL: EUROPE: Roy Edwards, *International Advertising Manager,* London; GWP, Düsseldorf. SEOUL: Biscom, Inc. TOKYO: Nikkei International Ltd.

ADMINISTRATION: John J. Moeling, Jr., *Publisher;* Marie D'Alessandro, *Business Manager*

SCIENTIFIC AMERICAN, INC.
415 Madison Avenue
New York, NY 10017
(212) 754-0550

PRESIDENT AND CHIEF EXECUTIVE OFFICER: Claus-Gerhard Firchow

CORPORATE OFFICERS: *Executive Vice President and Chief Financial Officer,* R. Vincent Barger; *Senior Vice President,* Linda Chaput; *Vice Presidents:* Jonathan Piel, John J. Moeling, Jr.

CHAIRMAN OF THE BOARD: Dr. Pierre Gerckens

CHAIRMAN EMERITUS: Gerard Piel

LETTERS

Cosmic Musings

To the Editors:

After reading "Universal Truths," by John Horgan [SCIENTIFIC AMERICAN, October 1990], I felt that I was hearing an updated medieval argument about how many angels could dance on the head of a pin. It seems that counting angels is no longer in fashion and that our philosophers, the cosmologists, now try to explain how the universe was created.

Is the universe infinite or finite? How does anyone postulate a universe with only an inside? And what about the big bang itself, the most bodacious explosion ever: a fireball a billion miles or more in diameter, expanding at light speed but maintaining a uniform density. Believing in that requires an act of faith equivalent to accepting that enumerable angels dance on pins.

By the way, it should be obvious I enjoyed the article.

ROBERT J. STARKS
Prospect Heights, Ill.

Whither the Dinosaurs?

To the Editors:

I was disappointed that the twin articles on the causes of dinosaur extinctions ["An Extraterrestrial Impact," by Walter Alvarez and Frank Asaro, and "A Volcanic Eruption," by Vincent E. Courtillot; SCIENTIFIC AMERICAN, October 1990] were not accompanied by a third. It could have explained our ignorance about whether dinosaurs disappeared gradually or suddenly, at different times in different places or all at once. We know too little about *what* happened 65 million years ago to say *why* it happened.

The global fires, super-chills, supergreenhouse effects, burning acid rain and similar catastrophes proposed by the advocates of impact and volcanism explanations are supposed to have been so severe that freshwater organisms were the main survivors. Yet many birds survived, too. Birds are so easily starved, poisoned and asphyxiated that they have become standard environmental indicators in coal mines and elsewhere. If many of them lived through the disaster at the end of the Cretaceous period (KT boundary), why not their dinosaur relatives?

After all, dinosaurs did survive the formation of enormous craters and volcanic traps in the Mesozoic era. Rather than dropping sharply in diversity and size after those events, dinosaurs steadily increased their diversity. Like large mammals, big dinosaurs had long limbs that allowed them to disperse widely. And unlike large mammals, which breed slowly, dinosaurs laid dozens or even hundreds of eggs. How could every breeding population of such animals be killed? The dinosaur extinction is one of the most extraordinary and complex events in earth history, and it is one that we will not understand for a long time.

GREGORY S. PAUL
Baltimore, Md.

To the Editors:

A recent analysis of 150,000 vertebrate fossils from the KT boundary that my colleague Laurie Bryant and I conducted shows that in the western interior of North America, only 35 percent or so of the species became demonstrably extinct. That figure, which is based on actual counts of specimens, differs from the 50 to 75 percent extinction figures given by the authors. As far as I know, those numbers are only estimates.

All the single-cause theories of extinction have a common weakness: they simply explain too much. The taxonomic groups that were most affected at the KT boundary were the sharks and their relatives, the lizards, the ornithischian and nonavian saurischian dinosaurs and the marsupials. Some of those declines can be tied to trends documented in the rock and fossil records, such as the fragmentation of habitats and competition between marsupial and placental mammals.

The fossil record is central to testing the idea of a KT mass extinction. Without a better understanding of it, all the asteroids and volcanoes are for naught.

J. DAVID ARCHIBALD
Department of Biology
San Diego State University

To the Editors:

Alvarez and Asaro show that the dinosaurs died out quickly, perhaps in as little as 50 years. Courtillot shows that the dinosaurs died out slowly. Each article seems to be scientifically bulletproof, yet at least one of them is full of holes.

NORMAN CUTTER
Las Vegas, Nev.

Editor's note:

Many readers wrote to suggest that an extraterrestrial impact might have triggered volcanic eruptions at the KT boundary. This possibility was briefly mentioned in both articles. No known geologic mechanisms, however, can yet couple the two phenomena.

Hemp Hassles

To the Editors:

I enjoyed your discussion of the many nonsmoking commercial uses of the hemp plant, alias marijuana ["Science and the Citizen," SCIENTIFIC AMERICAN, December 1990]. One small point needs to be clarified, however: I am the sole founder and national director of the Business Alliance for Commerce in Hemp (BACH). Jack Herer is not directly affiliated with us.

Ironically, in that same issue, "50 and 100 Years Ago" reprinted a 1940 item about how Du Pont's first nylon plant might replace a third of U.S. silk imports by 1942. Farm-grown hemp fiber could, of course, have met the entire silk demand—but it had been outlawed in 1938 at the urging of Du Pont and other special interests.

CHRIS CONRAD
National Director, BACH
Los Angeles, Calif.

ERRATA

An incorrect price was listed for a book reviewed in the January issue, *Rainforests: A Guide to Research and Tourist Facilities in Selected Tropical Forest Sites in Central and South America*, by James L. Castner. The price is actually $21.95, or $23.45 postpaid.

The table of contents for the February issue incorrectly listed the name of the third co-author of "X-ray Microscopes." The authors are Malcolm R. Howells, Janos Kirz and David Sayre. The accompanying paragraph also erred: three-dimensional imaging is theoretically possible, but it has not yet been realized.

In "Mathematical Recreations" for November 1990, the value of $10,585 after 5.5 percent inflation was misstated. It is $10,033.18.

We apologize for having suggested, in a recent profile of Robert C. Gallo ["Science and the Citizen," SCIENTIFIC AMERICAN, January], that Allan J. Erslev of Jefferson Medical College is deceased. He is still very much alive.

50 AND 100 YEARS AGO

APRIL 1941: "Several years ago The Universal Council for Psychic Research posted an award of $10,000.00 to any medium who can produce any supernatural manifestation which its Chairman, Dunninger, cannot duplicate or explain through scientific means. To this still standing award, Scientific American now adds $5,000.00 as a further incentive, the award to be available for two years."

"By means of language, tradition, and writing, the experiences of past generations can be handed on to present and future ones, and thus each generation may receive the knowledge accumulated in the past. Our knowledge is growing, but is our intellectual capacity increasing? Do the best minds of today excel the minds of Socrates, Plato and Aristotle?"

"Inside the Sun, the collisions between atomic nuclei cause a gradual transformation of hydrogen into helium, which supplies the energy that they radiate. These atomic encounters should be more numerous inside a white dwarf, in proportion to the density. With the same internal temperature we should expect thousands of times more heat to be produced than in the Sun, and yet the amount actually produced cannot be greater than that which the star radiates—less than 1 percent of the Sun's output. Here is a discrepancy of 1,000,000 to one, or more. The only way to explain it appears to be to assume that the atoms which, by their reactions, produce heat within the Sun, are enormously less abundant in the white dwarfs."

"Discovery that the musical ring can be squeezed out of copper and then pounded back into it suggests a clue to why some materials are hard and others soft. Dr. Thomas A. Read, Westinghouse Research Fellow, reported the discovery before the American Physical Society. He illustrated the discovery by gently tapping a bar of copper cast from molten metal. It rang as clearly as a tuning fork, but after it had been squeezed in a press, or even just dropped on a table, it answered with a dull 'clunk' when tapped with a hammer. Yet when the bar was pounded or cold-worked thoroughly, its lost ring returned. While no one is yet sure why the metal loses its ability to ring and then regains it, Dr. Read said, the most likely explanation is that dislocated copper atoms stifle the ring."

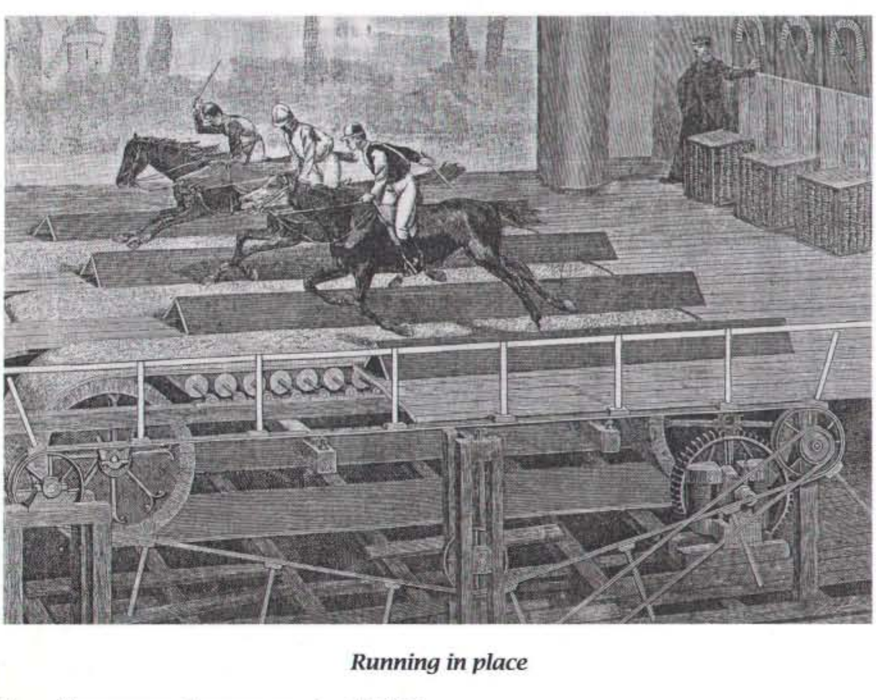
SCIENTIFIC AMERICAN

APRIL 1891: "As many years are now required as months formerly to build and arm a modern battle ship. What folly, therefore, to talk of creating a navy in an emergency. If we are to have a navy at all, let us have one that can whip the enemy if we must fight, and one that will be a school of the highest form of mechanical education if we shall be blessed with peace."

"Are we fatally obeying a law of nature—that law formulated in these words of the Scriptures: 'Dust thou art, and unto dust shalt thou return'? Dr. Variot, one of the most distinguished physicians of the Paris hospitals, answers this question by proposing to his contemporaries the use of electro-metallurgic processes for obtaining indestructible mummies. The doctor metallizes our entire cadaver. He surrounds it with an envelope of bronze, copper, nickel, gold or silver, according to the caprices or wealth of those who survive us. The inventor of the process accords to the total metallization of the body but slight importance. The object of his researches has been more especially to give the museums and laboratories of our faculties of medicine very faithful, very exact specimens, rather than to rescue our cadavers from the worms of the grave."

"A method of extracting teeth without pain was recently demonstrated in London. An electrical arrangement employs a couple of bichromate cells and a Ruhmkorff's coil to which is attached a commutator of extreme sensitiveness. A patient takes the handles of the battery in his hands. One of these is connected with the negative pole. The positive is divided into two, so that one of the divisions is connected with the handle and a wire from the other division is screwed into the handle of the tooth forceps. When the patient takes hold of the handles the current is gradually increased in intensity until the patient can bear no more; then, while the forceps are being introduced, the current is turned off for a second, and on again. 'Had you no pain?' asked our representative of the patient when the roots of the bicuspid were held up to view. 'Not a bit; I only felt the grip.'"

"One of the 'hits' of Messrs. Montreal and Blondeau's 'Paris Port de Mer,' played at the Varieties Theater, is a horse race. The horses are free from all restraint, and really gallop, but the ground disappears under their feet, moving in a direction opposite that of their running, and the landscape as well as the fences also fly in a direction contrary to the forward motion of the horses."



Running in place

SCIENCE AND THE CITIZEN

Cold Start

Policies on global warming and energy don't move critics

Genes for mortality, new neutrino conundrums, dioxin whitewash, John Sununu up close

When delegates from 101 nations gathered in Chantilly, Va., in February for the first round of negotiations toward an international agreement on global warming, they were presented with what appeared to be a new U.S. initiative. It was a slick color brochure detailing "America's Climate Change Strategy." There was also word that the Bush administration would soon roll out its long-delayed national energy policy.

To some present, it appeared that the administration might finally be moving toward an integrated strategy on global warming and energy use. But that first impression proved misleading. The climate change strategy offered nothing new. The U.S. energy policy, which was announced in late February, was a direct descendant of Reagan era supply-side economics that barely paid lip service to energy conservation.

In the end, the Intergovernmental Negotiating Committee had precious little to show for two weeks' work in Chantilly. Reined in by White House Chief of Staff John H. Sununu, U.S. negotiators agreed only to the establishment of two working groups: one to consider what kind of "appropriate commitments" might be made to counter warming and one to think about how they might be implemented. "They have determined only the shape of the negotiating table," complained Michael Oppenheimer of the Environmental Defense Fund.

It was clear before the conference began that progress would be difficult if the U.S., which emits 20 percent of the greenhouse gases that warm the atmosphere, continued to oppose cutbacks in emissions. Several European countries, as well as Australia and New Zealand, had previously committed themselves to reducing emissions of carbon dioxide by 20 percent over the next decade. Meanwhile Canada and Japan had said they will only stabilize their emissions. The U.S. document simply explained that actions the federal government has already taken will result in its greenhouse gas emissions in the year 2000 being "equal to or below the 1987 level." "No other country in the world, so far as I know, can make that claim," bragged President Bush's science adviser, D. Allan Bromley, to a Washington press conference.

The "strategy," however, is a prediction rather than a commitment, and it hardly represents a policy shift. The U.S. claim that its emissions in the year 2000 will be no greater than in 1987 is based on a new way of assessing greenhouse gas emissions that combines the future warming potential of many gases, not just carbon dioxide.

The gases listed in the new index of global warming potential include chlorofluorocarbons, which are already being phased out. Over the next decade, other domestic pollution and efficiency measures will probably reduce or slow the growth in emissions of other polluting gases, such as methane, nitrogen oxides, volatile organic compounds and carbon monoxide. Only because of these previously decided actions is the U.S. contribution to greenhouse warming expected to remain constant for the next 10 years—and then only if the economy performs as expected. Under present policies, total U.S. greenhouse gas emissions will start to grow again after the year 2000.

The U.S. negotiators took political advantage of the new counting method by successfully opposing moves that would have entailed separate examination of carbon dioxide sources and sinks. Advocates of more radical action to curb greenhouse emissions criticize the approach. Forty-one senators proposed a resolution taking the administration to task. One of them, Al Gore of Tennessee, who has made global warming a personal crusade, called the strategy a "dishonest subterfuge."

The go-slow approach of the U.S. negotiators was underscored by a report from the congressional Office of Technology Assessment that was issued during the Chantilly conference. The OTA takes issue with the notion that it is impossible to make substantial cuts in carbon dioxide emissions without sacrificing public comfort and convenience. Such cutbacks, Sununu has argued, reflect an "antigrowth agenda."

The OTA concludes that the U.S. could within the next 25 years decrease its emissions of carbon dioxide by 35 percent from 1987 levels without major technological breakthroughs. The cost of such a course is uncertain, although the OTA admits it could be as much as 1.8 percent of GNP. But a more moderate set of options that would limit increases to 15 percent over 25 years would probably save money, the OTA believes. If no special effort is made, in contrast, the OTA estimates that U.S. carbon dioxide emissions will increase 50 percent by the year 2015.

The technical options that the OTA identifies for reducing emissions include improving generating efficiency and encouraging conversion to more carbon-efficient fuels such as natural gas, as well as making greater use of non-

Results of an Energy Strategy

The Bush administration projects that its energy initiatives will have the following effects, based on a hypothetical "current policy base." These computerized projections, however, include changes already mandated by the 1990 amendments to the Clean Air Act.

INCREASE:

- Domestic oil production by 3.8 billion barrels a day to 12 million barrels a day by the year 2010.
- Nuclear power generation to 21 percent of installed capacity by 2030, from 20 percent at present, preventing a projected decline in share.
- Natural gas consumption by one trillion cubic feet, or 3 percent, by 2000.
- The share of electricity generated from renewable resources by 10 percent by 2010.

REDUCE:

- Oil imports by 3.4 million barrels a day by 2010.
- Total U.S. oil consumption from 22.5 million to 19.0 million barrels a day by 2010.
- Barrels of oil consumed for each $1 million of gross national product from 2.4 to 2.0 by 2010.

SOURCE: U.S. Department of Energy

fossil fuels. Also important would be increasing energy efficiency by the end-user, where the biggest savings would come from better building design and more efficient lighting and industrial processes. But in order to stimulate consumers and companies to make use of the most efficient and least polluting technologies, such measures as taxes, tradable permits, regulations, incentives and information programs would be needed.

These kinds of recommendations are glaringly absent from the administration's long-awaited energy policy, which carefully sidestepped the term "policy" with the designation "National Energy Strategy." Almost all the earlier proposed regulations and incentives for improved fuel efficiency and increased use of renewable energy were whittled out in cabinet and subsequent White House reviews.

Administration officials say such proposals were likened to Carter era legislation, which is the mark of political death in the antiregulatory White House. (Notably, the $2.5 billion of tax credits the oil industry received last year seem to be exempt from such criticism.) One proposal that never made it into the energy strategy was a plan to increase substantially fuel-efficiency standards for automobiles. Other casualties include mandatory building-efficiency standards and production incentives for renewable energy sources. Although the strategy calls for research aimed at improving fuel efficiency, no strong incentives are provided.

More than anything, the energy strategy reflected the Bush administration's foreign policy and a continuing faith in supply-side economics. Against the promise of a successful campaign in the Persian Gulf conflict, which sent oil prices plunging to near pre-war levels, the energy plan called for increased domestic exploration, particularly in the Arctic National Wildlife Refuge, and deregulation of natural gas pipelines and electric utilities. The administration contends that domestic energy supplies can be increased sufficiently to hold imports to 40 to 45 percent of demand through the year 2010. "Don't kid yourself, the war is about energy," declares Senator J. Bennett Johnston of Louisiana, who is part of a bipartisan coalition sponsoring legislation that would mandate conservation and efficiency.

The global warming debate is now going on behind closed doors, until the delegates meet again in June. Meanwhile the energy policy will be taken up by a Congress that appears keen to play a part in the debate. Some 80 bills have already been put forward that would make up for the measures deleted by the administration, including a measure that would increase automotive fuel efficiency to 40 miles per gallon over the next decade.

The debate is likely to be rancorous. As one disgruntled Energy Department official put it: "In order to decrease reliance on fossil fuels, you either have to raise the price of gasoline or regulate efficiency. What we have is neither." —*Tim Beardsley*

Tracking the Missing Carbon

One of the biggest uncertainties about the greenhouse effect is the mystery of the "missing carbon." Fossil-fuel burning adds about six billion tons of carbon to the atmosphere each year in the form of carbon dioxide, and deforestation and topsoil erosion may add three billion tons more. Yet the amount of extra carbon that appears in the atmosphere each year is only 3.5 billion tons. Another 1.5 billion tons dissolve in the ocean. The remaining four billion tons vanish without a trace.

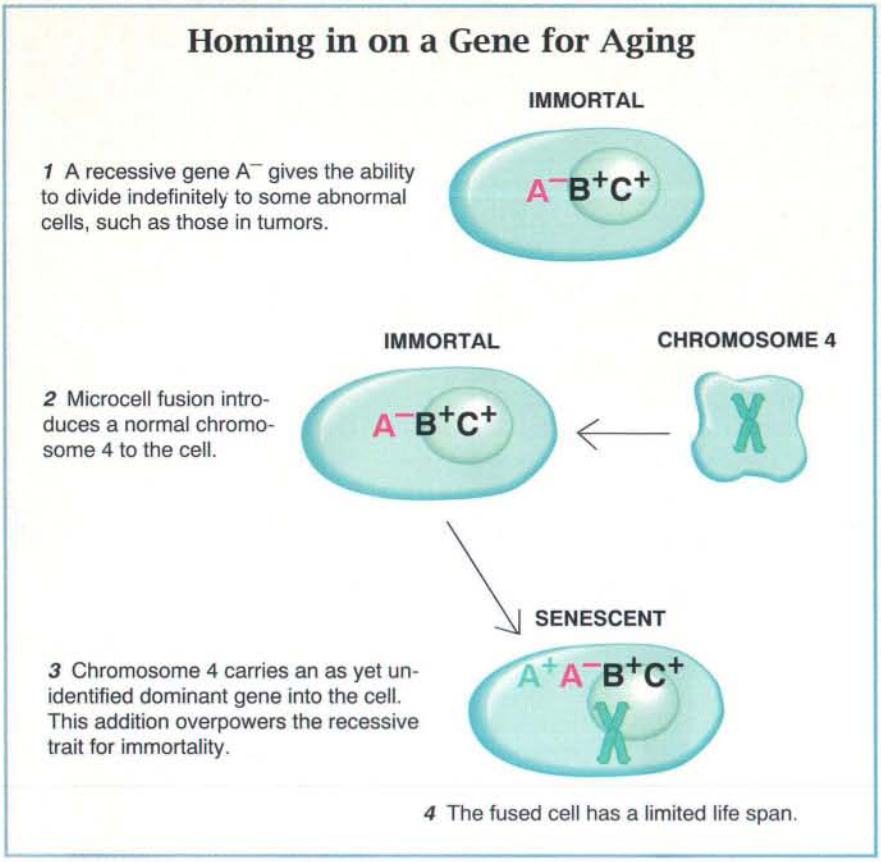
Some researchers think that trees may be a missing piece of the puzzle. They cite research showing that forests, stimulated by increasing levels of carbon dioxide, could soak up far more carbon than previously predicted.

Last year Pieter P. Tans, a researcher at the National Oceanic and Atmospheric Administration, and his colleagues concluded on the basis of computer models that two or three billion tons of carbon dioxide are disappearing into an unknown terrestrial "sink" that lies in northern temperate latitudes. Meanwhile experiments by Boyd Strain, a botanist at Duke University, and others have demonstrated a "fertilization effect" of extra carbon dioxide—plants grow faster and remove more of the gas from the atmosphere.

This effect, acting on forests, could in principle explain the fate of at least some of the missing carbon. Some experiments have shown that plants grow relatively larger root systems when carbon dioxide levels are increased. If trees are growing larger roots, researchers would probably not have noticed— weighing a tree's roots is a difficult operation.

To test the idea, Richard J. Norby, a botanist at Oak Ridge National Laboratory, is growing trees in atmospheres of elevated carbon dioxide with their roots in the ground (see photograph below). After two years, he reports that white oaks are showing sustained increased growth with extra carbon dioxide. White poplars respond marginally, but their leaves become smaller, a possible compensatory response. Norby has not yet dug up any trees to see whether their roots have grown faster.

Even if trees are storing missing carbon in their roots, the effect is not likely to affect greatly the pace of rising atmospheric carbon dioxide, says Anthony W. King of Oak Ridge. But Tans is more hopeful: "I can be an optimist and say, 'Great, nature is helping us here—we might be able to delay climatic change by growing forests on a large scale.'" —*Tim Beardsley*



Seeking Senescence

Specific genes may control how many times a cell can divide

Normal cells don't live forever. But some do continue dividing longer than others, seemingly controlled by a predetermined counting mechanism. Cells taken from humans or horses outlast those from mice, for instance. Cells from younger donors will double more often than those from older ones, as if they know they had more time allotted. The ability to divide slows as cells age.

One popular notion is that cellular aging is the cumulative result of small random changes, a breakdown provoked by a torrent of unintended insults. Others argue that senescence in normal cells is probably controlled by specific genes. "If you think that normal cells age because they are damaged, then immortal cells must have something that lets them escape," says Olivia M. Pereira-Smith, a geneticist at the Huffington Center on Aging at the Baylor College of Medicine in Houston.

Pereira-Smith and her colleague Yi Ning began examining cells that divide indefinitely, searching for a gene to induce aging and so stop proliferation. Their results, which will be published in the *Proceedings of the National Academy of Sciences,* indicate that one such gene involved in aging may be located on chromosome 4 in humans.

The Texas researchers first fused normal cells with immortal ones that were cancerous or virally infected, to see whether the hybrid offspring would also divide indefinitely. Most did not. The tendency to be normal—and thus of limited life span—is apparently the dominant inherited trait, whether cells originate in skin, muscle, veins or blood. Therefore, the few cells that do manage to become immortal must have a recessive gene for that trait, Pereira-Smith concluded.

Next the team began fusing immortal cells with other immortal cells, assuming that whatever recessive defect was responsible for each parent's longevity would be passed on to the hybrid. If parents shared the same genetic defect, the hybrid would also be immortal. But if parents were immortal and lacked the defect, their capacity for division would not be passed on to the hybrid. These parent cell lines would be assigned to different groups for indefinite division, depending on the other traits their hybrids inherited.

After fusing 30 cell lines from assorted tissues in all possible combinations, Pereira-Smith and her collaborators found that any given immortal cell line would fit into just one of four possible groups. This limited assignment indicates that a small number of highly specific genes are involved in senescence, she declares: "You need to turn on a certain set of genes to become senescent, so you have to lose any one of the set to become immortal."

Having firmly established which cells could be expected to live forever, Pereira-Smith decided to see whether they could be stopped from proliferating. "We decided to look at the chromosomes that have been implicated in tumor suppression," she notes, explaining that senescence could be involved in tumor suppression because it prevents cells from growing without control. Perhaps a gene sitting somewhere on one of those chromosomes would stop cells from dividing.

A procedure known as microcell fusion enabled the researchers to transfer a normal single human chromosome into an immortal cell line. The technique entails treating cells so that the nuclei break up into little bags of nuclear membrane, each one of which contains only one or two chromosomes. The microsacs are blown out of the cell cytoplasm and fused into the nucleus of another cell.

The first try was disappointing. When chromosome 11, which has been implicated in tumor suppression, was introduced to immortal cell lines, "they proliferated just fine," Pereira-Smith recalls. It may stop tumors but is apparently not involved in cell aging, she observes. The team chose to repeat the microcell experiment with chromosome 4. "We got lucky," the laboratory chief says with obvious delight. This chromosome was able to induce senescence in a line of cervical carcinoma cells but not in the other groups. Her laboratory continues examining other cell lines within the group to see if the effect can be replicated.

The next step is to find where on the chromosome the gene is located and figure out how it is regulated. Gene hunting is never an easy task, however. The classic example of the frustrating process is Huntington's disease. Researchers have known for more than 10 years that the gene for the inherited disorder lies somewhere on chromosome 4, but no one has yet been able to pinpoint it. But technology is getting better every day, the geneticist declares. Once the gene is in hand, scientists will be able to study its involvement in the senescence of various normal cells. Pereira-Smith muses that "maybe then we will be able to intervene not just in the laboratory but in aging people." —*Deborah Erickson*

A Gentler Therapy?

Retinoic acid turns off a form of leukemia

Like other forms of cancer, leukemia arises because of a failure in mechanisms that limit cell division. The result is an explosive proliferation of white blood cells. Because these cells are distributed throughout the body, the usual treatments, radiation and chemotherapy, are especially hard on the patient.

Now researchers in Shanghai, Paris and New York City have found a gentler way to curb at least one type of leukemia. Instead of killing the renegade cells in patients with acute promyelocytic leukemia (APL), they have succeeded in reprogramming the cells with a derivative of vitamin A, so that they cease dividing. The results are spurring redoubled efforts to determine whether the approach might be adapted to other cancers.

The principle, known as differentiation therapy, is based on the observation that cancer cells often seem to be stuck at an immature stage of development. Researchers have known for more than a decade that some chemicals stimulate immature cells to differentiate into their mature forms, whereupon they stop dividing. But attempts to turn the idea into a practical therapy have been disappointing.

One such group of differentiating chemicals is the retinoids, chemical cousins of vitamin A that have profound effects on cell differentiation. When retinoids were given to cancer patients in the early 1980s, there were some reports of beneficial effects but not enough to sustain interest. In 1988, however, Zhen-yi Wang and his co-workers at Shanghai Second Medical University treated APL patients with one retinoid, all-*trans*-retinoic acid. They reported that it brought about complete remissions in 24 patients.

Patients with APL normally have to be hospitalized because they are susceptible to massive hemorrhages that can kill in minutes. Even if they survive the initial crisis with the help of chemotherapy, two thirds die within five years. But the Chinese patients took retinoic acid at home in pill form, and its worst reported side effect was sore lips. "Frankly, the report was so staggering nobody believed it," recalls Raymond P. Warrell, Jr., a researcher at Memorial Sloan-Kettering Cancer Center.

One physician who was not so skeptical was Laurent Degos, a leukemia specialist at St. Louis Hospital in Paris. On hearing the trial results, Degos went to Shanghai and was so impressed that he procured supplies of the drug. At that time, it was produced in oral form only in China, even though it is the active ingredient in Retin-A, a skin cream marketed in the U.S. officially to treat acne. Despite the low level of purity of the Chinese drug, Degos obtained complete remissions in 24 out of 25 patients when he returned to Paris. Because of his success, the French subsidiary of Hoffman-La Roche, Produits Roche S.A., agreed to start producing an oral form of all-*trans*-retinoic acid in 1989.

When Degos told Warrell about his findings, Warrell decided to see whether he could achieve similar results. Fortunately, his wife, Loretta Itri, was then an assistant vice president of Hoffman-La Roche in the U.S. She urged the company to supply an oral formulation of all-*trans*-retinoic acid. The decision was not easy, because APL strikes fewer than 1,000 U.S. patients each year, too few to recoup a substantial investment. Also, the Food and Drug Administration was wary because the drug had no safety record.

Hoffman-La Roche finally agreed, and Warrell, in collaboration with Wilson H. Miller, Jr., Ethan Dmitrovsky and others, started U.S. trials last year. "Everything the Chinese and the French said is true," Warrell asserts. He and his team recently described remissions in nine of 10 APL patients and will soon be publishing results on more.

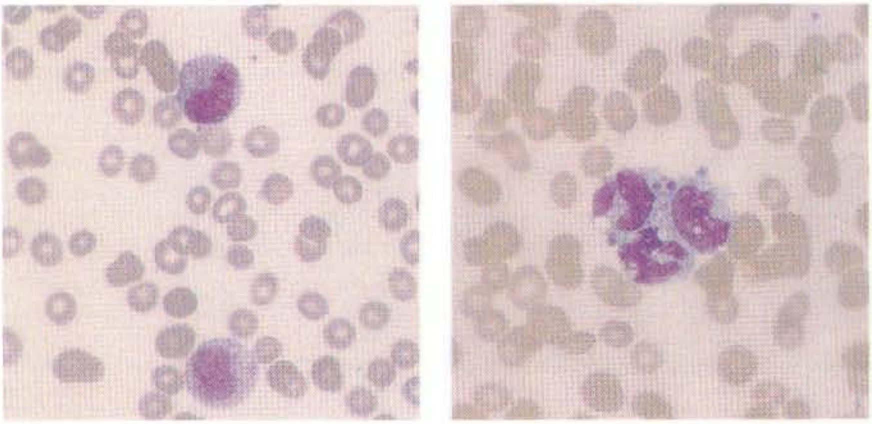
Retinoic acid, however, is not itself a cure for APL. If given the drug alone, most patients relapse after several months. But Degos says patients he has treated with retinoic acid followed by conventional chemotherapy have been disease free for more than a year.

Even so, the researchers believe the treatment is the most dramatic proof yet that differentiation therapy can work. "We can watch the cells go through the stages of normal maturation, ending up with an unusual but fairly mature" white cell, Warrell says. And Degos adds: "It is the first time we have brought about complete differentiation, changing the cells' programs to achieve normal cell death."

According to Dmitrovsky, the successful differentiation therapy "may serve as a model with application to other human malignancies," perhaps leading investigators to the control failures that cause them. Independently, both Degos and Dmitrovsky and their co-workers recently showed that most APL patients have an error at the site of a gene for a protein that binds retinoic acid in the cell. Moreover, one supposed APL patient who did not have the genetic error failed to respond to retinoic acid, suggesting that the error may be the fundamental cause of the disease.

In the test tube, retinoic acid causes other cancer cells to differentiate as well, including cells from teratocarcinomas—cancers of germ cells in the testes and ovaries—and some breast cancer cells. The National Cancer Institute is now conducting toxicity trials of Hoffman-La Roche's product, with encouraging results so far, according to Henry C. Stevenson, an NCI researcher. Stevenson says plans are being made to investigate the activity of retinoic acid against several kinds of cancer.

Buoyed by the apparent potency of all-*trans*-retinoic acid, Hoffman-La Roche and other drug companies are now searching for variants of it that might have other valuable properties. "Cytodifferentiation is an idea that has been anticipated for a long time," Dmitrovsky says. "To my mind, this actually works." —*Tim Beardsley*



LEUKEMIC CELLS circulate among smaller blood cells in a patient with acute promyelocytic leukemia* (left). *Treatment with all-*trans*-retinoic acid gives the leukemic cells more mature characteristics* (right). *Photo: Raymond P. Warrell, Jr.

Into the Lattice

Individual molecules reveal the dynamics inside a crystal

To study matter in its condensed states, scientists frequently rely on fluorescence spectroscopy, which analyzes the light emissions of excited molecules. Molecules have unique spectral signatures that reveal the properties of a substance, such as its chemical composition and bond lengths. These molecules, however, are a lot like relatives: they all show up at once. Traditional techniques "measure the average properties of thousands to millions of molecules," says W. E. Moerner of the IBM Almaden Research Center in San Jose, Calif.

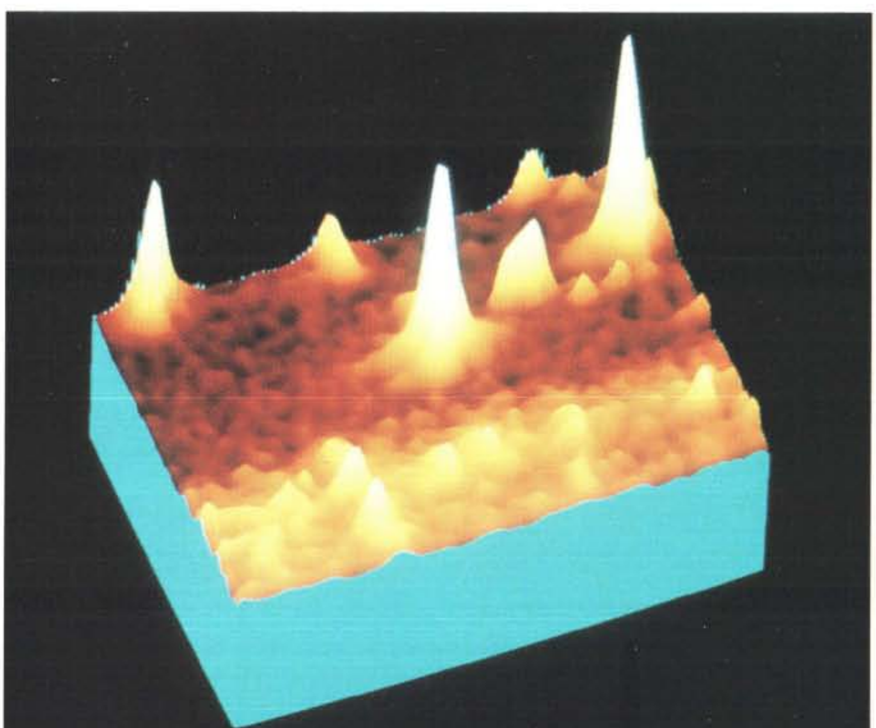
Consequently, the subtle movements of individual molecules in a crystal may become blurred in the big picture. To glean the missing information, Moerner and his colleague W. Pat Ambrose devised a way to analyze single molecules implanted in a host material.

Each impurity molecule acts as a probe, responding in distinct ways to the small deviations in the order of the host's crystalline structure. The ability to observe the changes in the impurity may prove important in the fundamental study of glasses, which as amorphous solids are highly disordered, and of semiconductors, which derive their properties from deliberately introduced impurities, or dopants. In principle, single-molecule detection techniques could form the basis for ultrasensitive molecular detectors and sensors and provide a means to observe biological transport with optically tagged molecules.

In their experiment, the researchers embedded molecules of the hydrocarbon pentacene in crystals of *para*-terphenyl, another kind of hydrocarbon. A pentacene molecule will fluoresce when excited by laser light, much like an empty bottle that resonates when blown across the top. A photon counter can measure the fluorescence.

Conceptually, the method is simple. But detecting single molecules in various environments has proved nettlesome. The dynamics of the host molecules obscure information from the probe molecule. Such crystal-lattice vibrations, called phonons, broaden the spectral lines of the probe molecules, making it difficult to distinguish individual molecules.

In order to tease information from the noisy background, Moerner and Ambrose used a micron-thin laser beam to probe crystals of *para*-terphenyl only one to 10 microns thick. Both the small laser spot and thin crystals helped to reduce the amount of scattered light. The researchers also cooled the samples to 1.5 kelvins so they could observe "inhomogeneously broadened lines" in the spectra. Inhomogeneous broadening results from imperfections in the host crystal; consequently, scientists can study the local environments, or the regions where the probe molecules reside.



INDIVIDUAL MOLECULES of pentacene in a host crystal emit light when excited by a laser. Each molecule shows up as a needlelike "bump" of fluorescence when plotted against its position in the host **(depth axis)** *and the laser frequency* **(horizontal axis).** *The intensity of each emission differs because the position of the molecules varies with respect to the focus of the laser.*

But narrow laser beams, thin crystals and frigid temperatures are not enough to investigate local environments. "There tends to be a bunching up of the molecules" around a favored frequency, Moerner says. So instead the researchers tuned the laser to the "wings" of the broadened line, or to the frequencies where few pentacene molecules would fluoresce.

Theoretical calculations helped Moerner and Ambrose to deduce that the resulting "bumps" in the measured spectra did indeed represent single molecules. Because the strength of the absorption and the amount of fluorescence are proportional to the number of molecules, the researchers knew they were dealing with one pentacene molecule and not multiple pentacenes emitting at the same frequency.

"But an even stronger proof that these are single molecules comes from the surprising spectral jumping effect we saw," Moerner explains. Ordinarily, when a solid cools to extremely low temperatures, all normal vibrations of the lattice cease. The pentacene molecules should therefore exhibit no variation in fluorescence.

Some of the molecules, however, did not remain fixed. Their resonance frequencies jumped in intervals ranging from one second to seven minutes. "That wasn't expected," Moerner observes. "There's motion of a new sort in the crystal." He believes that the jumping effect probably results from the nearby *para*-terphenyl molecules changing from one orientation to another.

Other probe molecules and host materials, such as glasses and polymers, can be studied with this kind of fluorescence spectroscopy, enabling workers to test theoretical models of condensed matter. So far, however, only probe molecules in the wings of the inhomogeneous lines—that is, those molecules isolated in frequency from most others—can be observed. Studying molecules where most bunch up is another matter. Still, Moerner thinks this spectroscopic technique may be able to provide the level of detail for molecules in solids that the scanning tunneling microscope does for atoms. "We are going to see unexpected things all the time for a while." —*Philip Yam*

Looking for Nothing

The taciturn neutrino keeps physicists guessing

Gertrude Stein once summed up Oakland, Calif., with the statement that "there is no there there." She might just as easily have been speaking about neutrinos. After all, it is hard to imagine anything that is not there quite so much as a neutrino, a particle that has no electric charge, that has little or no mass and that barely interacts with normal matter.

Nevertheless, neutrinos have aroused considerable scientific interest because they are extremely abundant throughout the universe and they travel freely from the center of the sun, the site of the nuclear reactions that ultimately sustain life on the earth. They may even provide a window into the true character of the four forces of nature. Yet recent studies have in some ways only underscored just how far scientists are from understanding fully the laws that govern neutrinos and the rest of the subatomic realm.

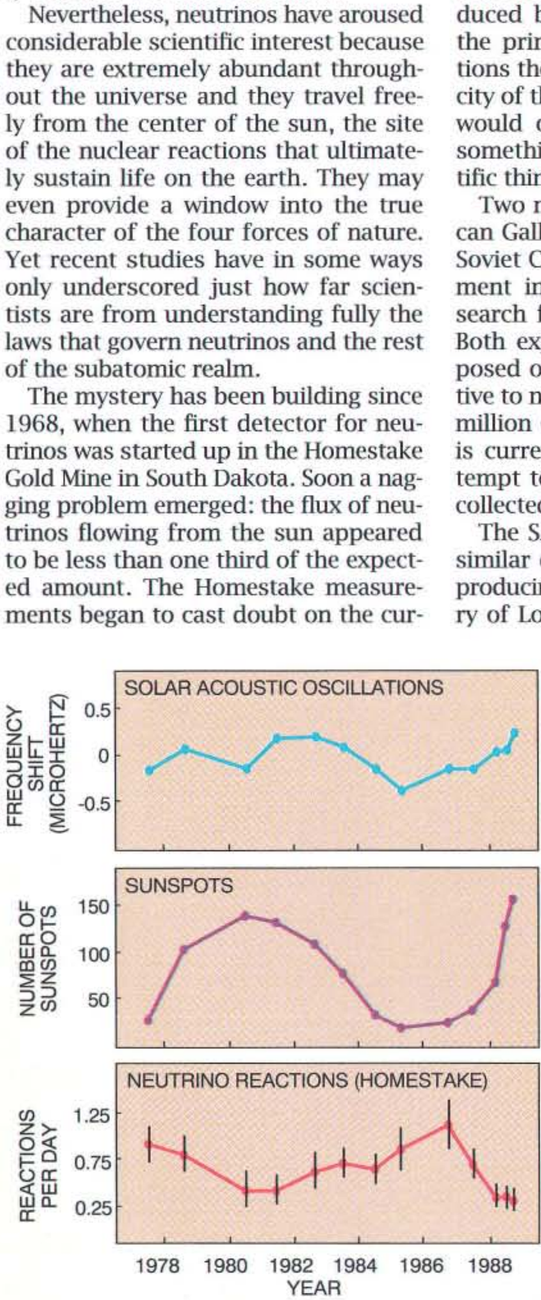
The mystery has been building since 1968, when the first detector for neutrinos was started up in the Homestake Gold Mine in South Dakota. Soon a nagging problem emerged: the flux of neutrinos flowing from the sun appeared to be less than one third of the expected amount. The Homestake measurements began to cast doubt on the current understanding of the laws of physics or of the manner in which the sun functions, or both.

A second detector, a Japanese experiment called Kamiokande II that joined the neutrino hunt in 1987, just intensified the confusion. This instrument, which is only sensitive to more energetic solar neutrinos (those having energies greater than about five million electron volts), found about one half of the expected flux, confirming the neutrino shortage.

Unfortunately, both the Homestake and Kamiokande detectors are blind to the relatively low energy neutrinos produced by the fusion of two protons, the primary link in the nuclear reactions thought to power the sun. A paucity of these "proton-proton" neutrinos would offer a far clearer signal that something is amiss with current scientific thinking.

Two new projects, the Soviet-American Gallium Experiment (SAGE) in the Soviet Caucasus and the Gallex experiment in Italy, have been initiated to search for these low-energy particles. Both experiments use detectors composed of gallium metal that are sensitive to neutrinos having as little as 0.23 million electron volt of energy. Gallex is currently on hold as physicists attempt to remove impurities that have collected in the detector.

The SAGE detector has completed a similar debugging process and is now producing its first results. Michael Cherry of Lousiana State University, a participant in the SAGE project, cautions that the margin of error is very high and that researchers need time to calibrate SAGE and improve their analysis techniques. Nevertheless, the early results have raised a few eyebrows. During its first five months of observations, SAGE has detected only background noise, but nary a sign of a solar neutrino.



SOLAR ACOUSTIC OSCILLATIONS and sunspot number appear to change inversely with the measured neutrino flux—depending on one's eyesight and on one's faith in the data.

Even the neutrinos that are showing up may be behaving in strange and unexpected ways. Some researchers, among them Kenneth Lande of the University of Pennsylvania, perceive a variation in the readings from the Homestake experiment: the flow of solar neutrinos seems to dip at times of high solar activity. In a related vein, Lawrence M. Krauss of Yale University finds hints of a correlation between the solar neutrino flux and solar acoustic oscillations. The implication is that the physical conditions that vary during a sunspot cycle—magnetic field strength or perhaps temperature and density—affect whatever mechanism hides most of the solar neturinos.

Then again, what looks like something might be nothing at all. Sunspot numbers have at times mirrored stock-market prices and the number of Beatles albums sold. John N. Bahcall of the Institute for Advanced Study in Princeton, N.J., a longtime stalwart of the neutrino hunt, thinks that correlations with the neutrino flux are similarly suspect. "They probably are not real, and they will go away," he says softly.

Bahcall notes that fluctuations do not appear in the Kamiokande measurements or in the first seven years of data from Homestake. But most of all, he sees no plausible explanation for what would cause such variations: "There is nothing in the data that says 'ah hah!' " as he puts it.

At least everyone agrees the central problem—the missing solar neutrinos—is for real. The most popular explanation for the neutrino shortage is the Mikheyev-Smirnov-Wolfenstein (MSW) effect. This slight modification of the latest group of physical theories—the grand unified theories—holds that under certain circumstances, the neutrinos associated with normal matter (called electron neutrinos) can transform into their even more evasive cousins, the muon neutrino and the tau neutrino.

Years of study have convinced most physicists that the mass of the electron neutrino is unmeasurably low. But a corollary to the MSW effect is that at least one kind of neutrino must have a mass. This possibility has excited many cosmologists because neutrinos could then provide much of the extra mass in the universe predicted by cosmological theories but not visible in astronomical observations.

Weighing neutrinos poses some practical difficulties, and so physicists have turned to theoretical models. Results from Homestake and Kamiokande hint that the neutrino shortage becomes more pronounced at lower energies. From this, Bahcall and Hans A. Bethe of Cornell University have inferred that if the electron neutrino is essentially massless (still the simplest assumption), the muon neutrino should have a tiny mass, about one thousandth of an electron volt. In comparison, an electron—the lightest particle now known—has a mass of 511,000 electron volts.

Such a light neutrino cannot satisfy cosmologists searching for missing

mass. Fortunately, they can turn to the third neutrino, the tau neutrino. Lande muses that one could extrapolate from Bahcall and Bethe's mass estimate to derive a mass of one to two electron volts for the tau neutrino, "which would place it at the lower edge of cosmological interest."

Masses more pleasing to cosmologists are also possible. Dennis Sciama of the International Centre for Theoretical Physics in Trieste recently proposed a tidy scheme to solve both the solar neutrino and the missing mass problems. In his scenario the heaviest neutrino would have a mass of about 29 electron volts and would eventually decay, emitting a burst of ultraviolet light. Arthur F. Davidsen of Johns Hopkins University reports, however, that an experiment on board the *Astro-1* spacecraft failed to find the distinctive radiation of such neutrino decays.

Then again, neutrinos might not be lightweights after all. Especially astonishing are recent reports of evidence for a neutrino weighing a hefty 17,000 electron volts based on studies of beta decay, wherein a radioactive atom emits an electron and a neutrino.

In 1985 John J. Simpson of the University of Guelph in Canada found that some of the time, the electrons appeared less energetic than theory predicted. He concluded that the missing energy was carried off in the form of mass by a neutrino. Eric B. Norman and his co-workers at Lawrence Berkeley Laboratory recently observed a similar effect in the decay of carbon 14 atoms, which he also attributes to a heavy neutrino, most likely a tau neutrino.

Bahcall considers the sighting of this massive neutrino "a serious issue." Beta decays could emit tau neutrinos only by violating a conservation law stating that the emission of an electron is always accompanied by an electron neutrino. Cosmologists would suddenly have too much mass on their hands unless the neutrino is unstable and decays within a few million years. And the existence of a heavy neutrino would confound physicists who hoped they were finally starting to find a coherent pattern in the masses of subatomic particles.

Norman and other researchers are collecting more data and studying the decays of other radioactive isotopes to confirm the reality of the heavy neutrino. He hopes that "in six months we will be able to settle this once and for all." Still, given the slippery nature of neutrino physics—and neutrino physicists—the larger issues are likely to remain a source of controversy for quite a while. —*Corey S. Powell*

Old Masters

Determining precisely when ancient artists limned on cave walls and cliffs figures of game and spirits as well as abstract symbols has long been a problem for archaeologists. Pictographs could be dated only using indirect evidence. Scientists could draw clues from the archaeological surroundings and matter deposited on the painted surfaces. They could also infer an approximate age from the style and content of the art.

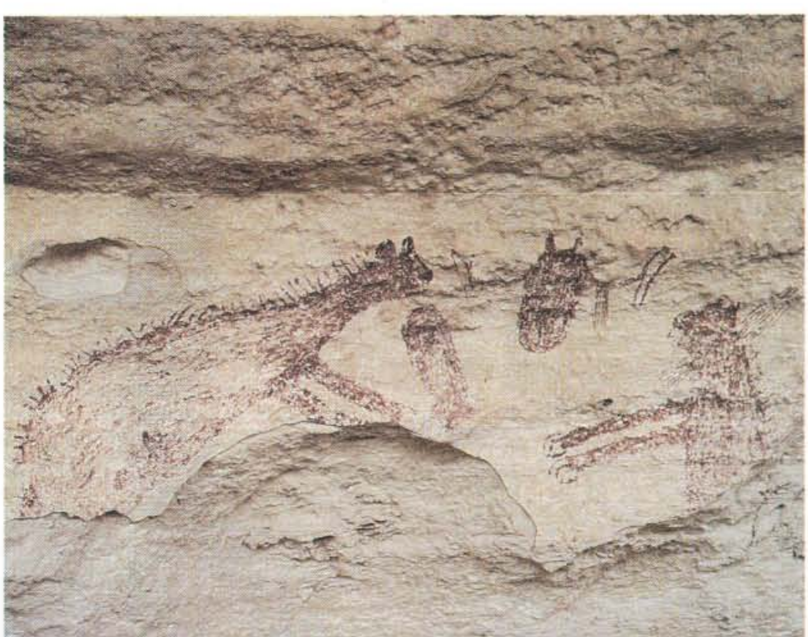
Chemists at Texas A&M University have developed a more precise technique. Marvin Rowe and his colleagues are using radiocarbon dating to determine the age of pictographs in southwest Texas. Their results, they say in *Nature*, prove that this method can be applied to pinpoint the age of any rock painting containing organic materials.

Radiocarbon dating, which is based on the decay of a naturally occurring carbon isotope, can determine the age of organic matter to within 100 years. Previously this technique was not widely used on pictographs, because it was difficult to distinguish carbon in the paints from that in the underlying rock. The researchers, however, found a way to separate the two types of carbon by exposing a tiny sample of the paint to an oxygen plasma. The highly energetic oxygen reacts selectively with the carbon in the paint. The resulting gas is then reduced to graphite, which is dated.

Harry J. Shafer, an anthropologist, first solicited Rowe's help in establishing absolute ages for cave paintings in the southwest region of Texas known as Lower Pecos (see photograph below). The paint pigments that were used in the Lower Pecos are primarily iron and manganese oxides, which contain no organic carbon. Yet Shafer knew that another group of paintings—in the Four Corners area of Utah—had been found to contain an organic binding agent, and he speculated that the prehistoric Texans used similar techniques to make their paint.

These ancient artists often used organic ingredients such as blood, eggs, seed oils, animal fats, milk, honey, plant resins and urine as binding agents. Such substances can be accurately dated. The Texas team successfully extracted organic carbon from a small piece of painting that had naturally peeled from a cave wall. A laboratory in Zurich dated the sample at 3,865 ± 100 B.P. (before the present). Shafer says this date correlates well with the date made by archaeological inference.

Shafer hopes to produce a more accurate chronology of ancient rock art by dating paintings characteristic of each period. "By using this method to date style differences in cave paintings," he says, "we can improve our understanding of the cultural sequence of the lives of ancient peoples." —*Nisa Geller*



History Lessons

Warfare analysts offer some disturbing—and hopeful—news

Political leaders always claim to be steering us by the lights of history toward a peaceful future. But what does a comprehensive analysis of our past actually reveal about our present course? A pessimist could conclude that our leaders are completely misreading—or misrepresenting—history. An optimist could find hope that warfare might become obsolete anyway—if the tentative spread of democracy worldwide continues.

These conclusions are both supported by the Correlates of War project, a computerized storehouse of information on 118 wars (defined as conflicts leading to at least 1,000 deaths) and more than 1,000 lesser disputes from the early 1800s to the present. Researchers at the University of Michigan created the data base in the 1970s to find statistical associations between warfare and various economic, political and social factors.

The data offer no support for the bromide "peace through strength," according to J. David Singer, a political scientist at Ann Arbor who oversees the Correlates project. A buildup of military armaments, far from deterring war, is one of the most frequent precursors of it. At the very least, Singer says, such a finding suggests that the U.S. policy of supplying arms to nations in an unstable region—such as the Middle East—is seriously flawed.

There is also no evidence that alliances help to keep the peace. In fact, a nation's participation in one or more alliances increases its risk of warfare, Singer says, particularly against its allies. History even casts doubt on the argument—used by the U.S. to justify both its current war against Iraq and its past one against Vietnam—that allowing aggression to proceed unchecked always leads to more aggression. Although Hitler's Europe certainly provides an important counterexample, Correlates of War data yielded little statistical correlation between warfare in a given region and prior unchecked aggression, Singer says.

A somewhat more hopeful finding

A Press Release on Dioxin Sets the Record Wrong

When the Chlorine Institute shopped around for a place to hold a scientific conference, they did not want just any host. "We were looking for an organization that was squeaky clean, that would not in any way, shape or form be questioned about the conference," says Robert G. Smerko, president of the Washington, D.C.-based institute, which is supported by some 170 chemical, paper and other manufacturers.

Smerko seemed to have met his requirements when he finally landed Cold Spring Harbor Laboratory. Last October the laboratory's respected Banbury Center held a conference—jointly sponsored by the Chlorine Institute and the Environmental Protection Agency—on the toxicity of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin, or TCDD. That chlorinated compound achieved notoriety during the Vietnam War, when it was identified as a contaminant of the defoliant Agent Orange. It remains controversial because it is found in some commercial herbicides and is produced in other chemical processes, such as paper bleaching.

Cold Spring Harbor Laboratory may have been squeaky clean, but the conference apparently was not. And the outcome of that meeting—attended by 38 of the world's dioxin experts, few of whom say they knew it was industry sponsored—is every bit as controversial as the substance that was the topic of discussion.

The issue is a press release sent out at the conclusion of the meeting by the Chlorine Institute's public relations firm, Daniel J. Edelman, Inc. It announced that the experts had agreed on a model for the toxicity of dioxin that "allows for the presence of a substance in the environment, with no risk experienced below a certain level of exposure." The release said that the scientists had rejected a linear exposure model, in which any level of exposure would have a biological effect, in favor of a receptor-based model that implies a threshold level. (This part of the release was approved by Cold Spring Harbor Laboratory, says the Banbury Center's director, Jan A. Witkowski—although he now says Edelman made several changes after he saw it.)

Such a consensus, of course, would have implications for setting permissible levels of the substance in the environment. But those at the conference insist that no such agreement was reached. "There was no consensus in terms of risk assessment," says George W. Lucier of the National Institute of Environmental Health Sciences. In addition, none of the scientists saw the press release, although their names accompanied it. "We were being used, clearly, and that's unfortunate," declares Arnold J. Schecter, professor of preventive medicine at the State University of New York at Binghamton. "Political layering is not particularly good, especially when it is unbeknownst," Lucier adds.

Few of the participants seem to dispute that the receptor-based mechanism of dioxin is relevant to human exposure. Nor did they before the conference, observes Alan P. Poland of the University of Wisconsin at Madison, who discovered the receptor in 1976. "The basic tenets were all known since 1981 or 1982," Poland says. But Lucier notes that now "we are at the point where we can reevaluate the linear model."

Indeed, the EPA intends to explore the question of whether there is a threshold response. The agency will investigate the receptor-based model with Michael A. Gallo, one of the conference organizers and a professor of toxicology at the University of Medicine and Dentistry of New Jersey–Robert Wood Johnson Medical School. But Gallo and others agree that discussion of thresholds in a regulatory context may be premature. At the conference, "some regulators got real excited by back-of-the-envelope calculations" and thought dioxin standards could be eased, says Linda S. Birnbaum, director of the EPA's environmental toxicology division. "Clearly, we don't know that."

Although many of the Banbury attendees were the last to know about the consensus they reportedly reached, news about the conference traveled quickly in political circles. At a recent hearing on dioxin standards in Alabama, expert witness for the pulp and paper industry Russell E. Keenan invoked the Banbury results in his testimony. "There was general agreement among the attending scientists that dioxin is much less toxic to humans than originally believed," Keenan claimed. Obviously, "it is not useless to tout Banbury results if you have a political ax to grind," comments Cate Jenkins, a chemist in the EPA's hazardous waste division. *—Marguerite Holloway*

is that democracies, although they often wage war against nondemocracies, rarely do so against other democracies. This phenomenon, which has held true among nation-states with virtually no exceptions since 1815, was pointed out in 1976 by Singer and Melvin Small, a historian at Wayne State University.

Since then, other researchers, and even Singer himself, have questioned the finding's significance. They noted, for example, that democracies tend to be wealthier than nondemocracies, that before World War II democracies were widely scattered and that since World War II many have been allied with the U.S. The lack of conflict between democracies, the critics argued, might arise from these factors more than from something intrinsic to democracy.

Yet a study completed last fall by political scientists Zeev Maoz of the University of Haifa in Israel and Bruce Russett of Yale University found examples of warfare among nations that met these other conditions but not the one of mutual democracy. Democracy, they concluded, seemed to be the crucial factor in maintaining peace.

The no-warfare effect also holds true among tribes and other less developed societies having participatory as opposed to hierarchical social structures, according to research by Russett and anthropologists Carol R. and Melvin Ember. The Embers oversee the Human Relations Area Files, a long-term study of warfare among societies that have not coalesced into a nation-state.

What explains the finding? Russett and the Embers speculate in a paper that "when people learn to agree to disagree and have some control over the political process, they may learn that conflicts can be resolved in peaceful ways, at least with people who share similar ideas about political process."

The proliferation of democracy in the past decade, although halting and by no means irreversible, has given Russett hope for a world without war. The growing economic interdependence of nations, he maintains, should provide an added incentive for peace. But he adds a big caveat: democratization may trigger violence by unleashing long-repressed nationalist and ethnic conflicts, particularly in Eastern Europe and unstable regions of the Third World.

Russett also warns that powerful democracies such as the U.S. should not use his findings to justify the violent "liberation" of supposed nondemocracies. One of the surest lessons of history, he notes, is that war often brings on not peace but chaos—and more war. *—John Horgan*

PROFILE: POLITICAL ENGINEER

John Sununu dominates science policy

When the elders of science used to bemoan the lack of science advice in Washington, they probably didn't anticipate that John Henry Sununu would fill the gap. Hear him on global warming, for example: "Frankly," says the White House chief of staff, "I have the feeling that a lot of people focus on carbon dioxide because what they are concerned about is not global warming, but their own antigrowth agenda. A lot of people who are moaning and groaning about global warming are also the same ones who are moaning and groaning about nuclear winter."

Before Sununu came to grips with the issue, President George Bush in 1989 endorsed the goal of stabilizing carbon dioxide emissions because they had been implicated in global warming. Now, as Bush's chief enforcer of domestic issues, the engineer who became the governor of New Hampshire has become the principal obstacle to that plan. "Frankly, I think he's hurting the president," says William A. Nitze, whom Sununu ordered fired from his position at the State Department because he openly favored controlling carbon dioxide emissions.

With a combative style that has earned him a reputation for delivering scathing put-downs to political allies as well as opponents, Sununu has become the administration's de facto architect of much of domestic policy. In a little more than two years in Washington, "the Governor," as he is known to his staff, has eclipsed key agency heads and advisers as the arbiter of issues from energy and technology to the environment—especially global warming.

"It's my sense there's a majority of opinion within the agencies that would like to see the administration go further in the direction of stabilizing emissions," says Alden Meyer of the Union of Concerned Scientists, the U.S. coordinator of Climate Action Network. "But it's impossible to have a rational discussion of the subject when Sununu is in the room."

Sununu, who has called himself a "political counterpuncher," has demonstrated his expertise at infighting in well-publicized battles with William K. Reilly, administrator of the Environmental Protection Agency, over wetlands policy and climatic change. With Sununu, together with Office of Management and Budget director Richard G. Darman and Michael J. Boskin, chairman of the Council of Economic Advisers, clearly in control, Secretary of Energy James D. Watkins seemed to sit on the sidelines as a national energy strategy was shaped by the White House. And even presidential science adviser D. Allan Bromley has been forced to backpedal on statements about greenhouse warming.

Yet it is Sununu's ability to absorb and retain details of technical issues that has allowed him to exert such a wide influence over policy. Born to Lebanese and Salvadoran parents in Havana in 1939, he says he knew that he wanted to be an engineer from about the age of seven, when he was given a book called *Engineers' Dreams.* He grew up in New York City and earned bachelor's and master's degrees from the Massachusetts Institute of Technology.

Sununu had more in mind than being a bench engineer, however. While still an undergraduate, he co-founded an engineering company, Astro Dynamics, Inc., and as its chief engineer designed heat sinks and brushless motors. He also worked on life-support systems for astronauts for the National Aeronautics and Space Administration.

But in 1965 he returned to M.I.T., where he earned a Ph.D. in fluid dynamics in only nine months. In 1966 Sununu was appointed an assistant professor of mechanical engineering at Tufts University, where he worked on problems involving heat transfer, temperature control and fluid dynamics.

Sununu began his move into politics soon after he and his wife, Nancy, moved to Salem, N.H., in 1969 to take advantage of that state's lower taxes. "Nancy and I really loved the state and decided that we ought to get involved to try to keep it the way we loved it," he says. So, even though he was commuting to Tufts, he found time to join the Salem local planning board. He soon became chairman and reveled in bringing his analytical acumen to political problems. Over the next 12 years, Sununu became increasingly involved in politics. He served a term in the state legislature and staunchly defended the controversial Seabrook nuclear power plant.

Although he had lost several political races, in 1983 he was elected the 93rd governor of New Hampshire. In three subsequent two-year terms, Sununu won a reputation as a fiscal conservative. Sununu's entry into the national arena came when he helped to secure

the president's New Hampshire primary victory in 1988. His reward was being made national co-chairman of the Bush campaign. He was tapped for the critical post of chief of staff soon after the election victory.

Now in the White House, Sununu puts in 13-hour workdays. And he clearly believes that his technical background is an important asset in Washington. In 1989 he told a meeting of the National Academy of Engineering that scientists and engineers—people who have a feel for problem solving and who "know the difference between a part per million and a part per billion"—have a public duty to become active in policy-making.

Nowhere has he more zealously applied his engineer's worldview to politics than on the issue of global warming. During the recent opening round of negotiations toward an international convention on climatic change in Chantilly, Va., Sununu and his staff kept U.S. negotiators on a tight leash, monitoring developments by telephone. Observers say it was Sununu who demanded that the word "appropriate" be inserted into the conference's final declaration on negotiating emission limits for carbon dioxide and other greenhouse gases.

Sununu insists he understands scientists' concerns on the question and is not opposed to lowering carbon dioxide emissions "as a matter of self-discipline." But he is adamant that the science of climate prediction is not developed well enough to take actions that might cause economic pain.

Yet his desire not to burden industry sits uncomfortably with the conclusions of the intergovernmental panel on climatic change, a group that has made perhaps the most authoritative assessment of the science of global warming to date. The panel concluded last year that without remedial measures, global average temperature was likely to rise one degree Celsius above its present value by 2025 and three degrees by the end of the next century.

Such changes would severely disrupt agriculture, natural ecosystems and human settlements. The National Academy of Sciences was scheduled to release a report in March that underscores these scientific concerns and argues that measures to slow emissions of greenhouse gases—such as improving energy efficiency—should be pursued now, even if they incur some short-term cost.

Sununu's basis for skepticism is simple: he doesn't trust the computer models used to predict climatic change. "What I have a problem with is the misrepresent—" Sununu stops and corrects himself. "—the gloom-and-doom approach to this that a lot of people have taken, based on some very preliminary analysis and modeling. They take models that are primarily two-dimensional across the surface of the earth and try to use them to characterize phenomena that are primarily driven by interactions in a vertical direction."

His distrust of models, Sununu says, comes from personal experience. (A persistent rumor maintains that he has a simple climate model that he runs on his personal computer. The models used by climatologists generally require supercomputers.) "I have a rule of thumb," he says, "that if you can't predict the past with a model, you ought not to believe you can predict the future."

JOHN SUNUNU believes those who "know the difference between a part per million and a part per billion" should enter politics. Photo: Brad Markel/Gamma-Liaison.

Sununu also says he finds it of "particular interest" that about 20 times more carbon dioxide comes from natural sources—primarily vegetative decay—than comes from burning fossil fuels. "The amount they are trying to get the international community to agree to not emit, if you will, over a short amount of time is less than the noise in the natural emissions. They ought not to be making decisions that affect the quality of life for billions of people based on something as marginally contributory as that."

But Sununu's criticisms of the science behind the concern about global warming do not gloss over the fact that he simply does not like to be told what to do. Commenting on the Chantilly conference, Sununu says: "Everyone seems to want to come over here and tell us what they would like us to do. We're saying, we've done a great deal, for a lot of different reasons. The net impact was positive—go home and do it."

When questioned about the administration's insistence on more research before taking deliberate steps to limit global warming, Sununu snaps that "we have put our money where our mouth is." He claims the U.S. has put $1 billion into climatic change research to ensure that "whatever we do is done intelligently."

Nevertheless, a well-publicized recent report written by Nobel laureate Leon M. Lederman, now president of the American Association for the Advancement of Science, argues that dramatic increases in support are necessary to avoid damaging the scientific research enterprise. Asked for his reaction, Sununu responds: "I don't know who Leon Lederman is." Instead he points out that the $76 billion requested in the 1992 federal budget for research and development—$8 billion more than this year—represents a "tremendous targeting of resources" aimed at nurturing the "fundamental strength we have in science and technology."

Sununu offers no apologies for his bluntness. "Any strong statements on my part are controlled, deliberate and designed to achieve an effect," he told journalists at the National Press Club last December.

Nor does the man who is so skeptical of greenhouse warming calculations hesitate to defend himself with statistics. "I have to interact with a member of Congress or a critical member of the administration or the press about 100 times a day. If I bat 99 percent in terms of having it be a very successful outcome all around, it means 365 times a year there's something that people can write a story about or exaggerate a story about." —*Tim Beardsley*

SCIENTIFIC AMERICAN April 1991 Volume 264 Number 4

The Real Cost of Energy

Bringing market prices in line with energy's hidden burdens will be one of the great challenges of the coming decades

by Harold M. Hubbard

In 1989 the U.S. Department of Defense spent more than $15 billion—as much as $54 billion according to some estimates—to safeguard oil supplies in the Persian Gulf. Early Pentagon estimates for the current war tack on another $30 billion. The smallest of these figures adds about $23.50 to the actual cost of each barrel imported into the U.S. By supporting this surcharge, the U.S. is subsidizing consumers of Middle East oil both at home and abroad.

This subsidy is only one of the more visible ways in which the U.S. and other nations affect the energy market. Energy costs society billions of dollars more than its users pay directly for oil, coal, gas or electricity. Other hidden costs of energy include tax credits, environmental degradation, increased health care expenditures and lost employment. Estimates for the U.S. alone range between $100 billion and $300 billion per year, although the precise amount is the subject of ongoing arguments among economists, environmentalists and policymakers.

HAROLD M. HUBBARD has spent most of his career working in energy and environmental research and development. He is now a visiting senior fellow at Resources for the Future in Washington, D.C., having recently retired as executive vice president of Midwest Research Institute. From 1981 through 1990 he was also director of the Solar Energy Research Institute. Hubbard received his Ph.D. in chemistry from the University of Kansas in 1951. He would like to acknowledge the assistance of Julie A. Phillips in the preparation of this article.

Daniel Sperling and Mark A. DeLuchi of the University of California at Davis have found, for example, that transportation fuels—gasoline and diesel—are priced well below their real costs. The 200 billion gallons of fuel burned each year for transportation constitute roughly a third of total energy consumption. Because vehicles are typically relatively inefficient and their emissions occur in proximity to people and buildings, they account for more than their share of energy's extra burdens.

In addition to the drag such buried costs place on the economy, they also distort the choices of both consumers and policymakers. Because the gap between sale price and actual cost is greatest for fossil fuels, unmasking these costs could provide strong incentives for a transition to more sustainable energy use.

Recognizing and minimizing the societal costs of energy—while still providing the energy required to fuel a growing global economy—will become an even more pressing concern in coming decades. The world's population is expected to increase from 5.5 billion today to more than seven billion by 2010. Moreover, demand for energy in the less developed countries will rise precipitously as they strive to increase their living standards.

The burdens that a barrel of oil or a kilowatt-hour of electricity imposes beyond its stated price are what economists call externalities: costs borne by people who are not parties to the transaction that imposes them. For more than two decades, environmental economists and ecologists have been struggling to identify and measure the external costs of energy production and consumption. Meanwhile conventional economics and current market policy ignore externalities, effectively setting their cost at zero.

Ignoring environmental and other social costs leads to what social scientist Garrett J. Hardin has called the "tragedy of the commons." Market forces lead inexorably to overuse of underpriced goods, be they public grazing lands, village dumps or free water supplies. Indeed, Hardin has charted the progress of civilization in terms of the internalization of costs formerly viewed as external.

Perhaps the first external cost to be internalized was that of raw materials. Even in prehistoric times, rules of private property and land ownership protected such valuable resources as copper ore—not to mention the fertile land itself.

Since then, lengthy and painful processes have internalized the cost of labor (by the abolition of serfdom), the cost of raising and educating the labor force (first by free public education and now in many countries by maternity leave and child care) and the cost of workplace safety (by workers' compensation and insurance benefits). Today chemical and thermal wastes generated by industrial processes—and energy

KUWAITI OIL TERMINAL burns after a naval battle in this wirephoto image from the Persian Gulf. Even in peacetime, oil imports necessitate billions of dollars in defense spending. The author argues that the market price of energy should better reflect its military, social, environmental and other hidden costs.

production in particular—pose a new challenge for internalization.

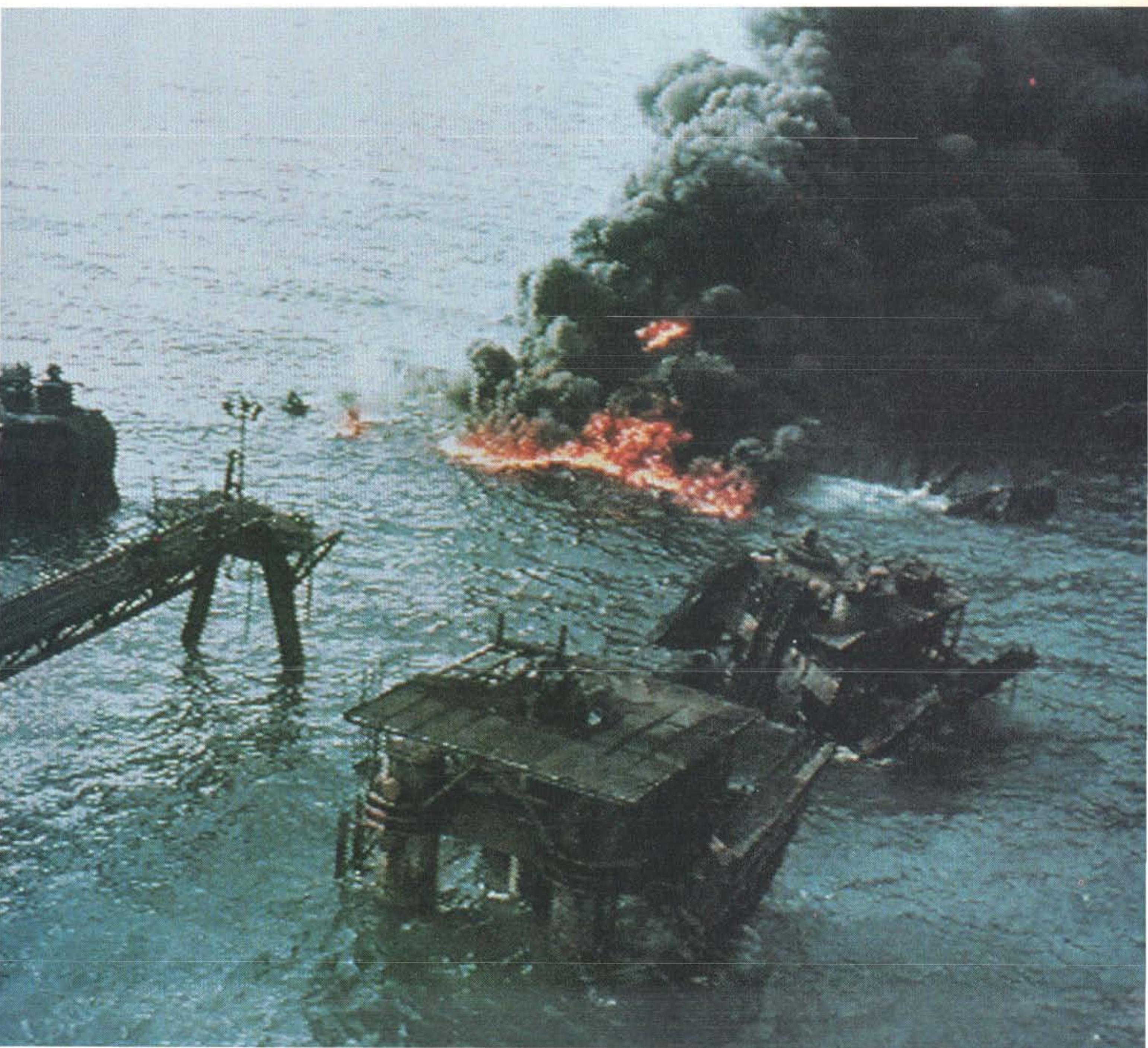
But calculating the actual cost of energy is not a simple matter. It is clear that consumption of different forms of energy generates costs beyond the market price, but the nature and amount of those costs are difficult to quantify. The answers that economists derive may depend as much on social values as they do on analytical solutions to well-defined problems.

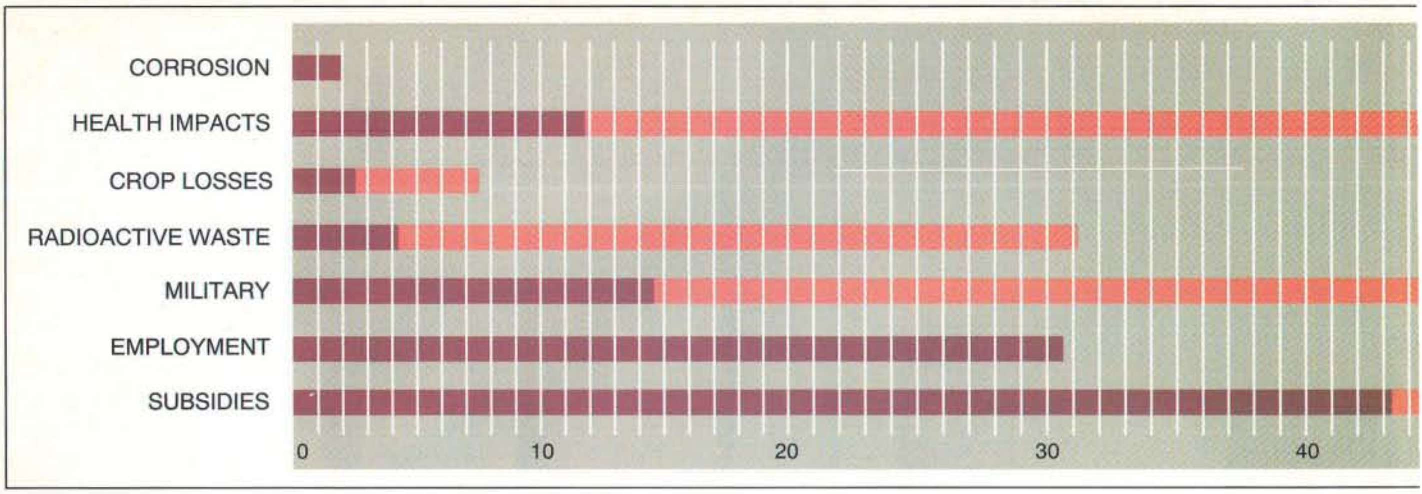
For example, the health effects of air pollution are an important added cost of both power generation and the consumption of fuels for transportation. Yet it is difficult to set exact prices for increased lung disease among senior citizens or lead poisoning among urban school children—let alone rank the two problems against each other. Although certain energy technologies impose a lower overall burden than others, all choices impose an unequal burden on some sector of the population.

Policymakers must also decide which external costs they are willing to take into account and which ones they will ignore. Consider the costs of energy-related accidents: Does one count only major disasters such as the Chernobyl explosion or the *Exxon Valdez* oil spill, or should every accident involving local gasoline and fuel-oil tank trucks be factored into the equation?

This question becomes even more difficult in the case of costs that are not easily quantified. Environmental costs include not only building corrosion, toxic waste emissions and crop losses but also generation of greenhouse gases, destruction of natural habitats and loss of species diversity. How are these to be valued? Similarly, how does one gauge the costs of the dislocations caused by the boom-and-bust cycle of oil prices, and who should pay it?

Economists are making progress toward answering some of these questions. A number of methods can be used to calculate various portions of the real cost of energy. Among the simplest to calculate are the subsidies—roughly $50 billion per year—that the federal government delivers to energy producers in the form of tax





Estimates of societal burdens created by energy consumption and production vary widely depending on researchers' initial assumptions. The largest variation is in health effects, which reflect both the difficulty of putting a price on human life and disagreements as to what part of pollution-related disease costs should be ascribed to energy use.

credits and research funding. About $26 billion goes to fossil fuels, which supply 85 percent of the roughly 80 quadrillion BTUs consumed in the U.S. each year. Another $19 billion subsidizes nuclear power, which supplies about 7 percent of U.S. energy consumption, and $5 billion supports renewable energy sources, which supply about 8 percent.

Researchers have used three disparate methods to calculate the environmental effects of energy consumption: control cost, damage cost and contingent evaluation. Control cost is the simplest measure and also the least accurate: it merely tallies the amount a power-plant operator, automobile manufacturer or other enterprise must spend to comply with existing environmental standards. Alternative technologies can be compared according to their control costs.

Unfortunately, the expense of meeting federal regulations bears little or no relation to the damages imposed on society by a given pollutant. Permissible emission levels (and thus the cost of controlling them) typically depend on political compromises among interest groups rather than technical understanding of the problems at hand.

Furthermore, calculations of control costs are usually specific to particular pollutants and control technologies; they may not address all the risks involved, and they may not even represent the most cost-effective approach to solving a particular environmental problem. (For example, scrubbers used to remove sulfur emissions from coal cost significantly more than just burning low-sulfur coal to start with.)

Measuring actual damage to the environment yields figures that could be much more appropriate for choosing among energy options. This method, however, requires a detailed understanding of how each of the pollutants emitted by a given energy source moves through the atmosphere; how it undergoes chemical transformations; how it is deposited in water, on lung surfaces and elsewhere; and how much it affects each entity it encounters.

In principle it should be relatively easy to estimate the damage done by various energy sources to market-valued goods, such as crops or building materials. In practice, analysts often disagree. For example, the Environmental Protection Agency estimates that energy-related ozone emissions reduce U.S. crop yields by 12 percent, costing more than $2.5 billion per year. Researchers at Cornell University have placed the lost yield as high as 30 percent. Similarly, it has been estimated that if the U.S. produced all its own oil rather than importing it, the increased employment of U.S. workers would be worth $30 billion to the economy as a whole. Many economists, however, would dispute this conclusion, asserting that workers who lose energy-related jobs are absorbed into other sectors of the economy.

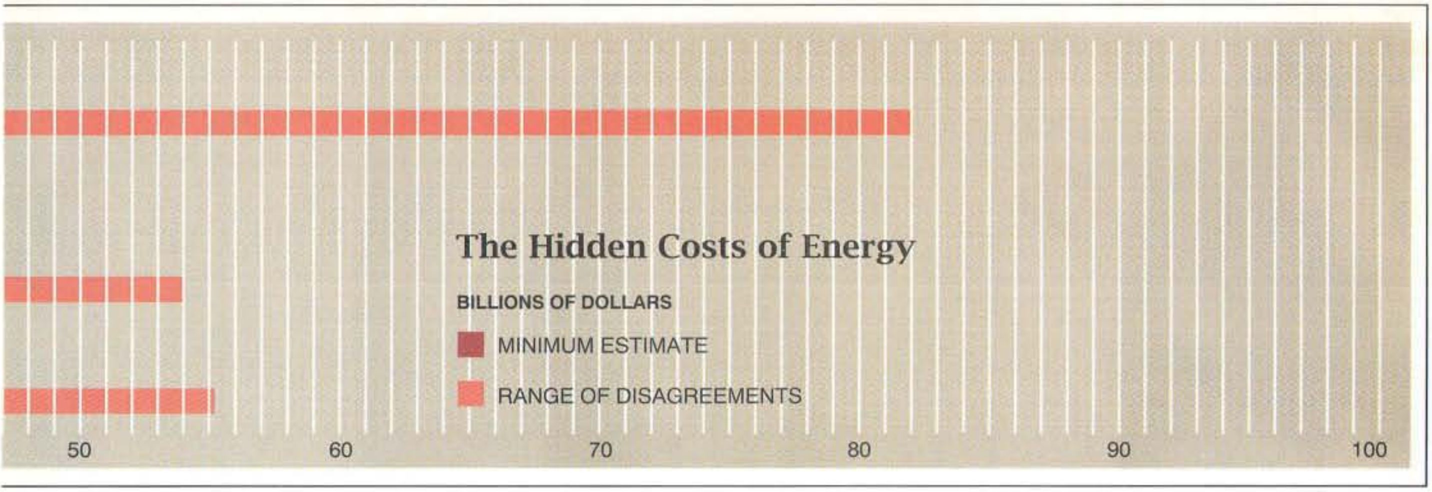
Moreover, many of the entities that suffer from environmental degradation are not commercially traded, and so their value is not easily measured. Among them are historical monuments, species diversity, ecosystem preservation, wildlife and wilderness areas, and visibility.

Insisting on a precise value for such items is simply a recipe for prolonged inaction. Researchers can set a reasonable price by indirect methods, such as contingent evaluation. Surveys can find out how much people would be willing to pay to avoid incurring a specific kind of environmental damage or other social cost—or, conversely, how much compensation they would ask to be willing to suffer a particular insult to their surroundings.

In 1990 the Bonneville Power Administration canvassed rate payers in Oregon, Washington, Idaho and Montana to determine how much they might be willing to pay for alternatives such as solar energy, wind power or increased energy conservation. A representative sampling of these 3.6 million consumers indicated that they would be willing to pay $13.30 each, or a total of $48 million extra a year, to avoid having a nuclear power plant built near them. Consumers were willing to pay $46 million to avoid a fossil-fuel plant and $20 million to avoid a new hydroelectric plant. These are hardly insubstantial sums to apply to the development of alternative energy technologies.

Contingent evaluation can also put prices on potential future environmental risks, such as nuclear accidents or global warming, whose likelihood and cost are both unquantifiable by conventional means. Although such valuations must be treated with caution, they may furnish a middle ground for action in situations where policymakers might otherwise be caught between advocates of intense anticipatory crisis management and those who insist on a wait-and-see approach.

During 1989, the Office of Conservation and Renewable Resources at the U.S. Department of Energy conducted two preliminary studies on the external cost of electric power



generation. The studies compared air pollution, land use, material requirements and carbon dioxide emissions from fossil-fuel, nuclear and renewable energy technologies at all stages of generation from initial fuel extraction to decommissioning. The results provide a mechanism for comparing the environmental impacts of energy technologies that have very different characteristics.

One important fact highlighted by the DOE studies is that no energy technology is completely environmentally benign. For example, although photovoltaic cells emit no pollutants during operation, their manufacture requires large quantities of hazardous materials, and their ultimate disposal could release toxic elements such as arsenic and cadmium to the environment. The studies found, however, that complaints of excessive land use leveled at solar and renewable energy technologies are inaccurate: producing energy from coal or uranium has similar real estate demands. The only difference is that coal, uranium and biomass

ACID RAIN damage is among the best-established consequences of energy consumption. Its cost is difficult to assess, however, because it affects assets such as forests, streams and lakes, whose market value is not readily quantified.

consume land at fuel-extraction sites, whereas solar and wind power consume land at the generating site.

Some of the choice among various energy technologies, then, is a matter of selecting the kind and location of externalities. Nevertheless, the DOE workers did find an overall difference in total externalities. Nuclear power (under normal operation), photovoltaics and renewable energy sources clearly do less damage to their environment than does coal burning.

The DOE policy office has commissioned further studies at Oak Ridge National Laboratory and Resources for the Future to analyze all the net social costs of different methods of energy production. For each fuel cycle under study—fossil, uranium or renewable—workers will measure all resource impacts, including requirements for labor, capital and materials as well as energy's effects on air and water quality and on national security. This project is proceeding in conjunction with similar studies by the European Community and other nations. When the work is completed in 1992, policymakers will be better able to compare existing and emerging energy technologies.

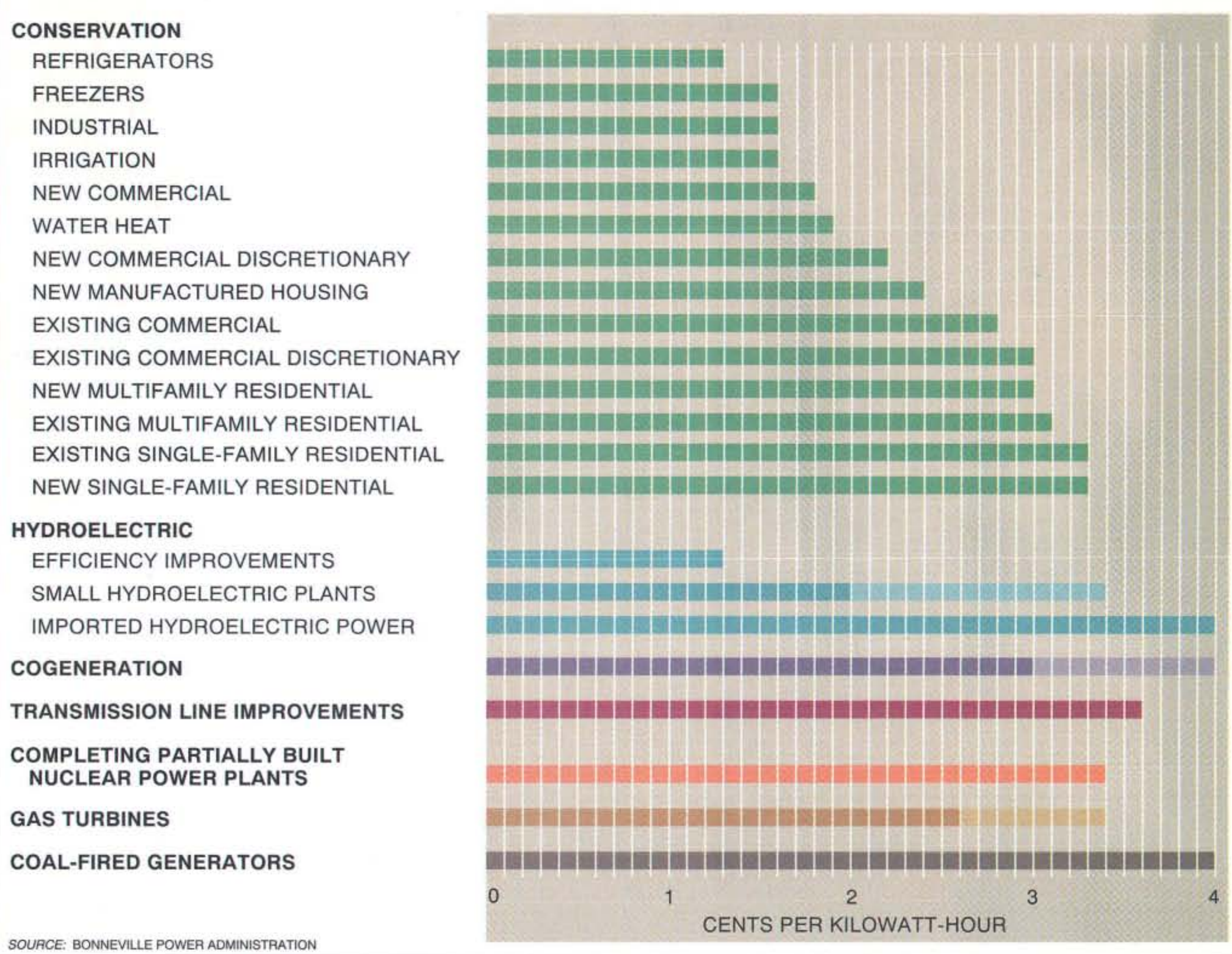
How should the external costs of energy be internalized? Taxes on energy offer a simple means to harness the actual power of the marketplace. A number of economists have estimated, for example, that a tax of $28 per ton on carbon-containing fuels would stabilize emissions of carbon dioxide, the primary cause of the greenhouse effect, over the coming decade. It would also reduce emissions of other pollutants.

Taxes may not always be the best method for internalizing energy costs. In some cases, such as electricity generation, the choice of energy source is irrevocably made long before anyone starts paying taxes on the amount of fuel consumed. Many state utility boards have turned to a technique known as integrated resource planning, which aims to meet energy demand at the least possible overall cost.

This method entails examining the costs of many different options in addition to simply building new power plants. Factors considered include conservation, efficiency improvements in existing systems, conversion to new fuel sources and nontraditional resources such as wind or solar energy. So far 26 states have begun the process of requiring public utilities to account for external costs in their integrated resource plans; 18 have already mandated that this be done.

As a first step, public utility commis-

The Cost of Meeting Electricity Demand



SOURCE: BONNEVILLE POWER ADMINISTRATION

Making additional electricity available by investing in efficient buildings and equipment is less expensive than building new power plants. This analysis includes both direct costs and social burdens. (Lighter shades indicate a range of estimates.)

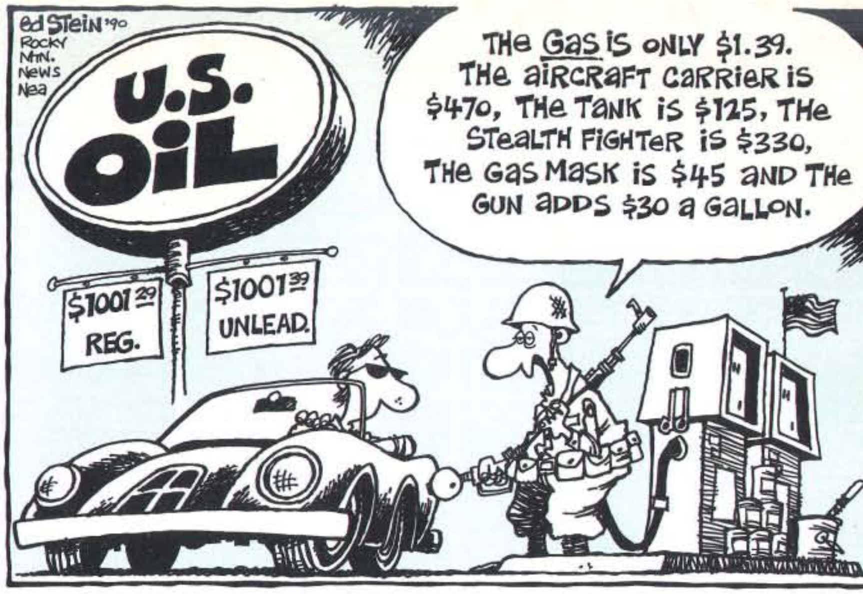
sions often decide to give preference to technologies that provide economic and environmental benefits without specifying exactly how the preference will work. The next stage is to include externalities by adding a fixed percentage to the estimated costs of polluting technologies (or by subtracting a similar percentage from the estimated costs of conservation or alternative energy technologies). These so-called percentage adders are a simple, although arbitrary, way of bringing environmental costs to bear in the marketplace.

Wisconsin, for example, considers either conservation or nonfossil fuel-generating capacity economically comparable to a conventional system even if it costs as much as 15 percent more. The state's public service commission also requires utilities to include the future cost of mitigating air-quality problems in proposals to renovate existing power plants. These estimates must take into account the increased expenditures that will be needed to meet future emission standards, which are expected to be more stringent than those in effect today.

Several states have begun employing more sophisticated methods that value environmental costs directly. New York State, for example, adds a certain amount per kilowatt-hour based on air and water pollution and land degradation to the expected price of electricity from each source. A coal-burning power plant without scrubbers, located in the middle of a city, would merit a penalty of 1.405 cents per kilowatt-hour, or nearly 25 percent of its total cost. Other power sources are penalized proportionately less.

Yet other states, including California, Oregon, Wisconsin, New Jersey and Colorado, are examining even more complex methods, such as those developed by Shepard C. Buchanan and his colleagues at the Bonneville Power Administration. These approaches are based on the total cost per kilowatt-hour produced or saved over the lifetime of a power-generating or conservation resource. The total cost includes the capital cost of building a plant or purchasing conservation equipment, the incremental costs of generating or saving each kilowatt-hour and various identifiable environmental and social costs. Conservation receives an additional 10 percent bonus because of its many unquantified environmental benefits.

Applying this scheme to the Bonneville Power Administration's own needs for the next 20 years produced surprising results. The most cost-effective resources were conservation and efficiency improvements in existing hydroelectric facilities. The most expensive options—new nuclear plants, cogeneration and coal-fired generators—cost two to three times as much per generated kilowatt-hour.



Although such a study considers conservation as a way of meeting electricity demand, as regulations stand today, virtually the only way for utilities to increase profits is to sell more electricity. Regulations must be systematically revised to allow utilities to make a profit from conservation by their customers.

Work by the DOE and others is making an important contribution to understanding the societal costs of energy. Translated into energy-pricing mechanisms, this understanding will allow the marketplace to move toward wise and profitable long-term energy-use decisions. Currently, however, the federal government is lagging while enlightened states such as California, New York and Wisconsin are leading the way, particularly with respect to planning for electric power generation. The federal government should follow suit by imposing taxes on fuels to match their external costs.

Government can help in other ways: by establishing performance standards that encourage the development of new products that save energy, by educating the public about the need for a broader perspective on energy pricing and by funding research and development of advanced energy technologies.

Sooner or later the public has to pay the real cost of energy. Paying those costs sooner means higher prices for gasoline, heating oil and electricity. This will be uncomfortable, even painful, for consumers. Attempting to pay later, however, may cost consumers far more and require more drastic action.

If the coming decades bring unprecedented climate changes and an increasingly poisonous atmosphere, consequences may include severe curtailment of heavy industry, banning of private automobiles in urban centers, mandatory restrictions of heating and air conditioning, and escalating global conflicts over resources and pollution.

Clearly, the sooner that policymakers and citizens act prudently to internalize all the costs of energy, the better. Paying societal and environmental costs up front will mean higher prices for today's energy consumers. But these higher prices will encourage more efficient use of energy and will favor technologies that are cost-effective for society as a whole. When prices reflect full costs, the market is an excellent system for strengthening the economy and for dealing with such serious issues as clean air, global climatic change and energy security.

FURTHER READING

BLUEPRINT FOR A GREEN ECONOMY. David Pearce, Anil Markandya and Edward B. Barbier. London Environmental Economics Centre, Earthscan Publications, 1989.

SOCIETAL COSTS OF ENERGY: A ROUNDTABLE. Boulder, Colo., American Solar Energy Society, 1989.

THE ENVIRONMENTAL COSTS OF ENERGY SUPPLY: A FRAMEWORK FOR ESTIMATION. Alan Krupnick. Resources for the Future, 1990.

ENVIRONMENTAL ECONOMICS: A SURVEY. Maureen L. Cropper and Wallace E. Oates. Resources for the Future, 1990.

The Structure of Quasicrystals

Quasicrystals are neither uniformly ordered like crystals nor amorphous like glasses. Many features of quasicrystals can be explained, but their atomic structure remains a mystery

by Peter W. Stephens and Alan I. Goldman

When aluminum, copper and iron are melted together and cooled, they can solidify to form a grain in the shape of a perfect dodecahedron, a geometric solid whose 12 faces are regular pentagons. Although this dodecahedral grain looks like a crystal, it is not. Crystals are composed of identical building blocks called unit cells, each containing precisely the same distribution of atoms and each fitting together with its neighbors in the same way. A dodecahedral grain cannot be constructed from atoms in unit cells of a single shape whether they be small cubes or even dodecahedrons. The dodecahedral grain is a quasicrystal.

Indeed, all probes of atomic-scale structure show that quasicrystals are not made up of repeated unit cells. It is clear that these exotic new materials cannot be crystals, but it is not immediately apparent just what they are. As physicists, chemists and materials scientists have investigated the structure of quasicrystals, they have come to realize that periodic crystals, whose atomic structures they have studied over the past 78 years, are but a subset of the possible types of ordered solids.

Since 1984, when Dan S. Shechtman and his colleagues at the National Institute of Standards and Technology (NIST) discovered the first quasicrystal, workers have fabricated many different species of quasicrystalline alloys [see "Quasicrystals," by David R. Nelson; SCIENTIFIC AMERICAN, August 1986]. They have learned how to improve samples to the point that they can quantitatively study issues that were only idle speculations seven years ago. The study of quasicrystals has advanced three theories about their structure: the Penrose, glass and random-tiling models.

The Penrose model—derived from the work of mathematician Roger Penrose of the University of Oxford—suggests that quasicrystals are composed of two or more unit cells that fit together according to specific rules. The model accurately describes some of the basic properties of quasicrystals, but it has difficulty explaining how these rules might be related to atomic growth processes.

The glass model, in contrast, relies on local interactions to join clusters of atoms in a somewhat random way. According to the model, all the clusters have the same orientation, but because of random growth, the structure contains many defects.

It now seems that the two models are converging toward a third, the random-tiling model, which combines some of the best features of its predecessors. In the past few years the structure of quasicrystals has been one of the most hotly debated topics in solid state physics. The resolution of this debate may lead to a theory of quasicrystalline structure and guide the development of materials with unusual structural and electrical properties.

To produce the first quasicrystals, Shechtman and his colleagues at NIST melted together aluminum and manganese and then squirted the molten metals against a rapidly spinning wheel, thereby achieving a cooling rate of about one million kelvins per second. This abrupt cooling process, called quenching, can "shock" the alloy into a variety of novel structures, or phases. To understand these unusual phases of matter, one must first have a grasp of some of the principles of basic crystallography.

A crystal can possess only certain symmetries because there are a limited number of ways that identical unit cells can be assembled to make a solid. For instance, a salt crystal is composed of cubic unit cells that stack to form cubic grains. Consequently, the salt crystal has fourfold rotational symmetry: when the crystal is simply rotated through a quarter turn around the appropriate axis, atoms of the rotated crystal occupy the same positions as those of the unrotated crystal. Crystals

PETER W. STEPHENS and ALAN I. GOLDMAN have collaborated on many projects during the past 10 years and now share a keen interest in quasicrystals. Stephens is associate professor of physics at the State University of New York at Stony Brook. He recently spent a year at Tohoku University in Japan to conduct experiments, some of which are described in this article. Goldman is associate professor of physics at Iowa State University and is a physicist at Ames Laboratory. After receiving his Ph.D. from Stony Brook in 1984, he worked at Brookhaven National Laboratory for four years.

QUASICRYSTAL (*right*) can be distinguished from other forms by its unusual symmetry. The material can form grains in the shape of dodecahedrons, solids whose 12 faces are regular pentagons. The dodecahedral grain has fivefold symmetry, that is, when rotated by one fifth of a circle about an axis through one of the faces, its appearance is unchanged. The grain is composed of aluminum, copper and iron and is about 300 microns in diameter. The electron micrograph (*above*), which shows a slice through several atomic layers, also exhibits fivefold symmetry.

can have only twofold, threefold, fourfold and sixfold symmetry.

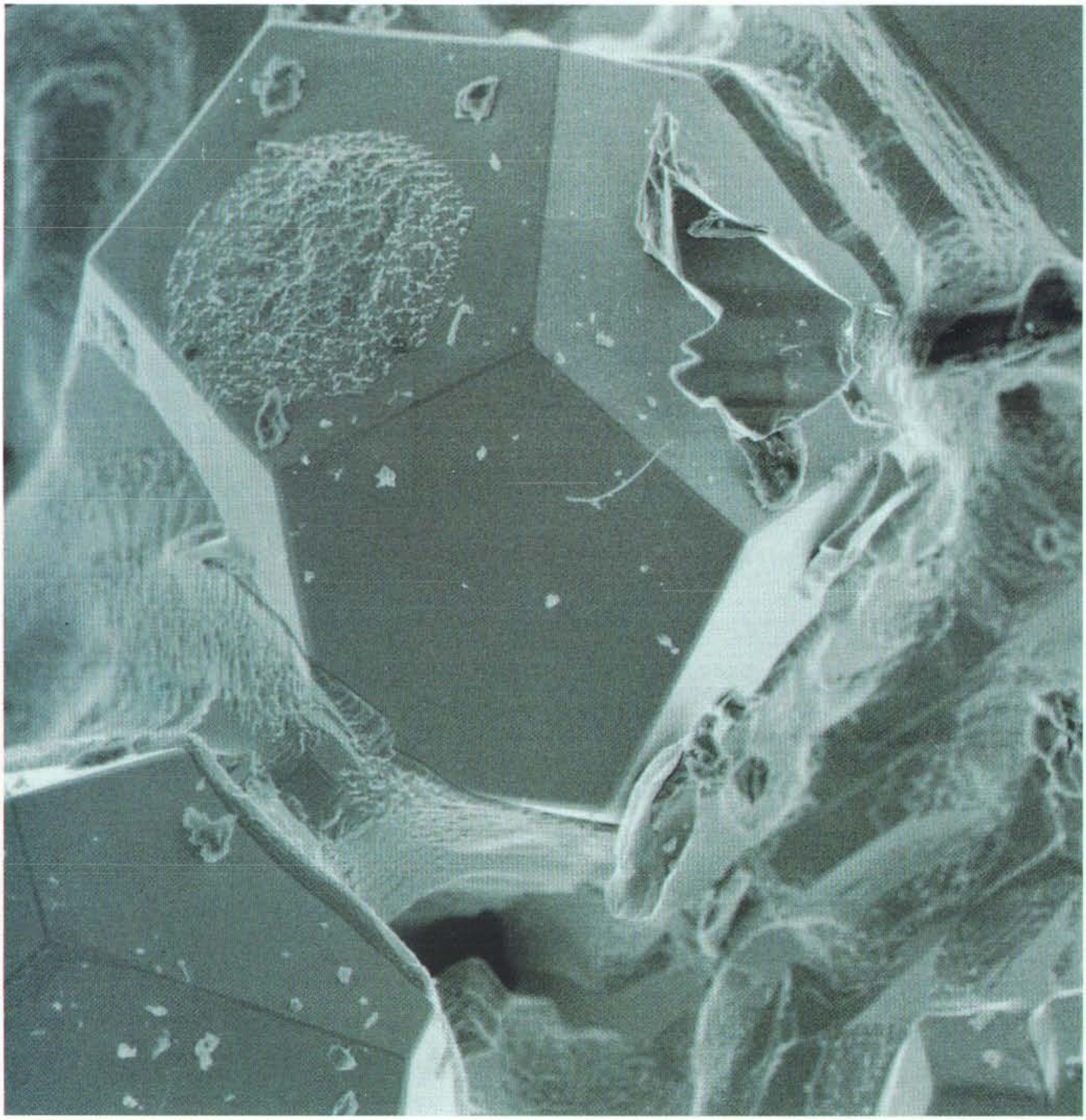
A crystal can never have, say, fivefold symmetry, because a single unit cell that has fivefold symmetry, such as a dodecahedron, cannot be assembled to completely fill space. There will always be gaps between the dodecahedral unit cells.

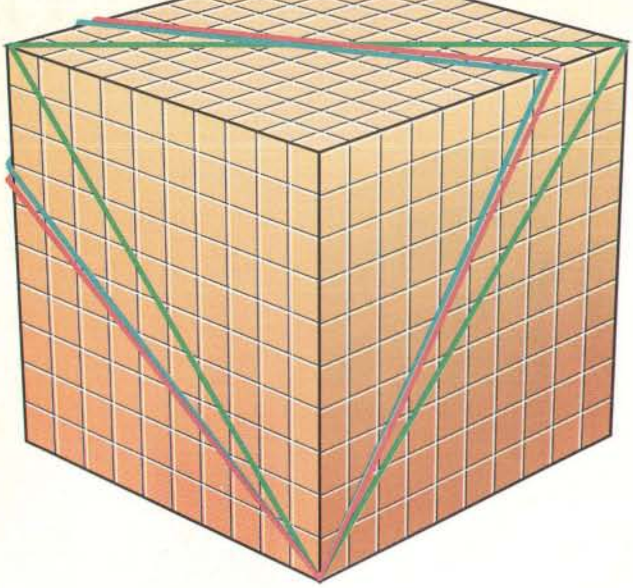
To determine the structure of a crystal, investigators use an indirect, but well understood, technique. Atoms in a crystal are arranged in families of parallel planes. Each such plane acts as a mirror to incoming X rays, electrons and other rays or particles that travel through space as a wave. Each plane reflects the incident waves very weakly. But if the reflected waves from each member of a family of planes combine in phase, the total intensity of the reflected wave can become quite strong. This phenomenon is called diffraction; it occurs whenever any type of wave interacts with an ordered structure of the appropriate spacing.

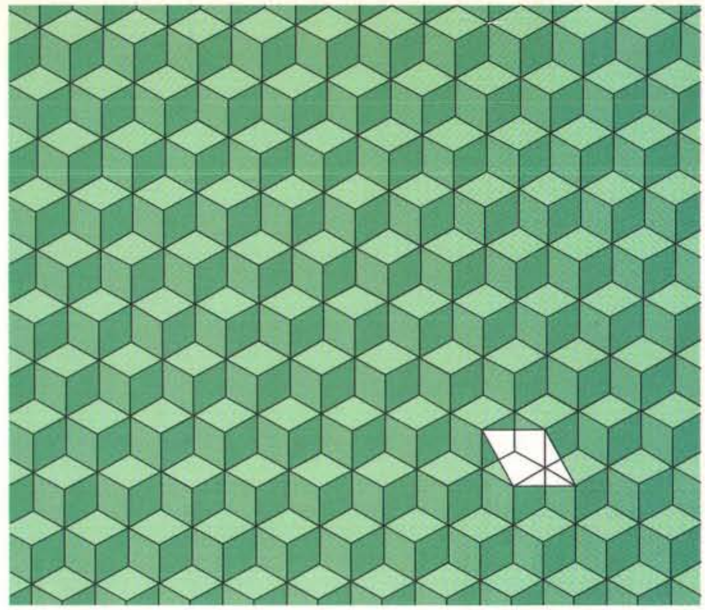
When a crystal is bombarded by a beam of X rays or electrons, the angles through which the waves are diffracted reveal the shape and dimensions of the unit cells of the crystal. The diffracted waves can be recorded, for example, on a photographic emulsion, where they appear as a pattern of bright spots.

All the symmetries of a crystal are reproduced in its diffraction pattern. A crystal with sixfold rotational symmetry will produce a diffraction pattern that also has sixfold symmetry. Because no crystal can have a fivefold symmetry axis, one would not expect to see a diffraction pattern that has fivefold rotational symmetry.

Nevertheless, when Shechtman and his colleagues illuminated a grain of the aluminum-manganese alloy with electrons, they found a diffraction pattern that had fivefold rotational symmetry.







PERIODIC OR QUASIPERIODIC pattern can arise when a plane cuts through a stack of cubes (*left*). The green plane intersects the cubes to form the green pattern at the right. It is periodic in the sense that the pattern can be assembled by duplicating and positioning a unit cell (*white area*) in an orderly fashion. The blue pattern is also periodic and has a unit

Indeed, by rotating the sample by the appropriate angles, they deduced that the alloy had six fivefold symmetry axes. In technical terms, the alloy has icosahedral symmetry, because an icosahedron is a 20-sided solid that has six fivefold rotational axes in the same orientation as the alloy.

The icosahedral alloys are only one of many families of quasicrystals discovered during the past seven years. All these materials have symmetries that are "forbidden" in conventional crystals. Leonid A. Bendersky of NIST found that aluminum and manganese can form a material that is periodic along one direction and has tenfold rotational symmetry in the perpendicular plane. Workers have also recently fabricated quasicrystals with eightfold and 12-fold symmetry. In some sense, these quasicrystals provide a link between quasicrystalline and crystalline order. They also demonstrate that the phenomenon of quasicrystallinity extends far beyond ideas about icosahedral symmetry and the stability of specific icosahedral clusters of atoms.

The first model for quasicrystals emerged from the mathematics of tiling—a field advanced by Penrose and others during the 1970s [see "Mathematical Games," by Martin Gardner; SCIENTIFIC AMERICAN, January 1977]. Penrose examined how two or more shapes could be assembled in a quasiperiodic way to tile a plane, that is, to cover it completely with shapes that do not overlap. Each of these quasiperiodic tilings, now known as Penrose tilings, could be constructed using a set of instructions called matching rules [*see box on page 29*].

In 1982 Alan L. Mackay of the University of London calculated the diffraction properties of a theoretical quasiperiodic structure. He demonstrated that if atoms were placed at the corners of each shape in a Penrose tiling, they would give rise to a diffraction pattern that had tenfold symmetry. Then, in 1984, Peter Kramer and Reinhardt Neri of the University of Tübingen extrapolated the concept of two-dimensional Penrose tilings to three dimensions. Dov Levine and Paul J. Steinhardt of the University of Pennsylvania had also been considering the possibility of alternative forms of atomic order based on Penrose tilings.

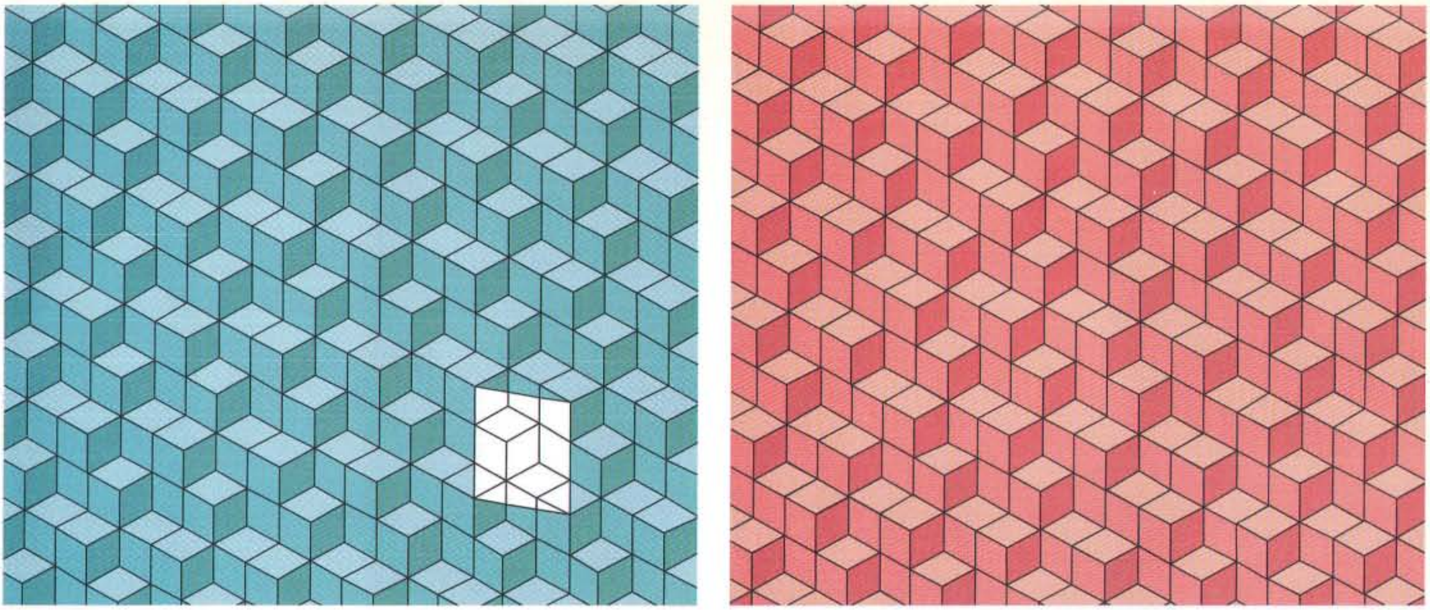
Building on this early work, Levine and Steinhardt proposed the Penrose model for quasicrystals only six weeks after Shechtman and his colleagues published their famous paper. They put forth a three-dimensional generalization of the Penrose tiling that described the structure of the aluminum-manganese alloy, and they showed that the pattern of diffraction peaks calculated from their model agreed well with Shechtman's results.

Penrose quasicrystals are constructed from a set of unit cells and specific matching rules that govern how they fit together. These rules are more complicated than the identical repetition of identical unit cells that form a crystal. Three important features distinguish a Penrose quasicrystal from a crystal.

First, a Penrose quasicrystal contains many regions that explicitly show forbidden rotational symmetries, that is, fivefold. Second, a Penrose quasicrystal is built from two or more unit cells rather than a single unit cell as suffices for periodic crystals. Third, a Penrose quasicrystal does not exhibit equally spaced rows of lattice points as the periodic structures do. Nevertheless, the diffraction pattern resulting from a Penrose quasicrystal has an array of sharp spots, in agreement with the experimental observations.

One can elegantly describe the structure of Penrose quasicrystals and their corresponding diffraction patterns if one thinks of Penrose structures as resulting from a slice through a higher-dimensional periodic lattice. This concept is easiest to visualize in two dimensions.

Imagine a two-dimensional lattice composed of points that sit at the corners of squares in a grid. A horizontal line of these points is covered by a strip. If the covered points are projected onto a line that is parallel to the strip, the projected points will be equally spaced along the line. Those points define a periodic sequence because they divide the line into equal-size segments. To produce a quasiperiodic sequence, the strip must be tilted with respect to the lattice so that it has a slope equal to an irrational number [*see illustration on page 28*]. (A number is irrational if, when expressed as a decimal, no set of consecutive digits repeats itself indefi-



cell. The red pattern, which cannot be constructed from a single unit cell, is quasiperiodic. Whether a pattern is periodic or quasiperiodic depends on the angle of the cut. Whereas the slope of the green plane with respect to the stack of cubes and the slope of the blue plane are equal to rational numbers, the slope of the red plane is an irrational number.

nitely, for example, the number π, or 3.1415....) If the points that are covered by the strip are then projected onto a line parallel to the strip, the projected points will divide the line into a quasiperiodic sequence of long and short segments.

This sequence serves as a one-dimensional model of a quasicrystal if one imagines that an atom is placed at each point that divides the line into long and short segments. Like other quasicrystals, the quasiperiodic sequence has unusual diffraction properties. One might guess that the one-dimensional model would generate a blurry diffraction pattern because the atoms are not periodically spaced. But careful calculation proves otherwise. In fact, the quasiperiodic sequence produces a sharp diffraction pattern as a consequence of periodicity of the two-dimensional parent lattice.

The diffraction pattern derived from a quasiperiodic sequence consists of a dense set of weak and strong peaks [*see illustration on next page*]. In experiments, one detects only the strongest of these peaks. Even so, one can still show that the quasiperiodic sequence is not a crystal because of the aperiodic spacing between the observable peaks.

To make quasiperiodic structures in two dimensions, one needs a lattice in a space of even higher dimension. For example, a plane that cuts through a stack of cubes at an angle can form a quasiperiodic two-dimensional structure if the slopes between the plane and the cube axes are irrational numbers [*see illustration above*]. (Indeed, a two-dimensional Penrose tiling that has fivefold symmetry can be produced by a projection from a five-dimensional lattice.)

Using similar reasoning, theorists have described icosahedral quasicrystals as a three-dimensional cut through an abstract six-dimensional lattice. This kind of cut leads to the type of quasiperiodic structures proposed in the Penrose model. By describing quasicrystals in this way, one can understand how quasicrystals can have unusual symmetries and yet not contradict the precepts of crystallography. Whereas the possible symmetries of the diffraction patterns of crystals are limited by the possible symmetries of the crystals themselves, the symmetries of quasicrystals derive from those of a higher-dimensional parent lattice.

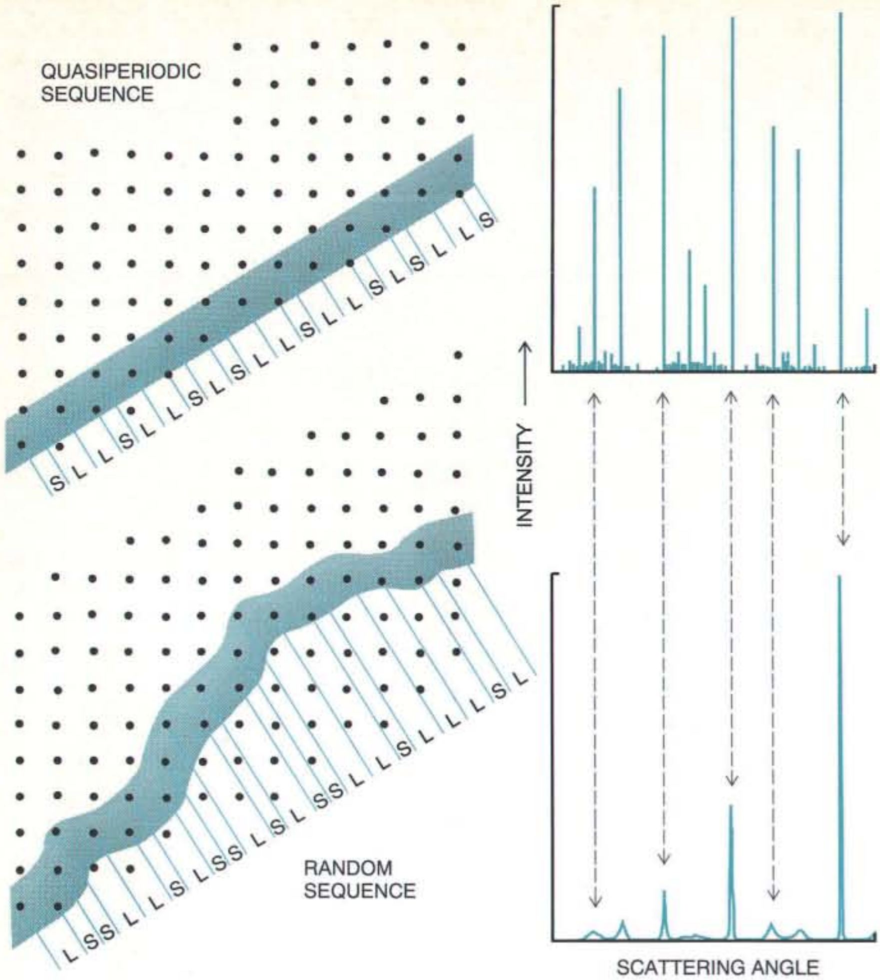
In general, a quasiperiodic structure will give rise to a diffraction pattern that has great order because of the periodic order in the higher-dimensional parent lattice.

Although the Penrose model is very successful in predicting the diffraction patterns generated by icosahedral alloys, it gives few clues about how physical reality is related to multidimensional spaces and matching rules. In particular, although the Penrose matching rules are local, a great deal of planning ahead is required to construct a perfect Penrose quasicrystal. Growing such a quasicrystal requires that atoms in very distant unit cells would have to interact in some manner to communicate their positions and relative orientations. This idea is contrary to all generally accepted notions about crystal-binding forces, which are relatively short range.

Another objection to the Penrose model is that it fails to account for the considerable disorder evident in almost all quasicrystals. This disorder appears in their structural, electrical and diffraction properties. For instance, one consequence of the Penrose model is that a perfect quasicrystal should conduct electricity as well as an ordinary metallic crystal. In fact, all quasicrystals produced in the laboratory have conducted electricity rather poorly.

A more important sign of disorder is revealed in the X-ray diffraction patterns from icosahedral alloys. In many cases, these show broadened peaks in contrast to the perfectly sharp peaks predicted by the Penrose model. Broadened diffraction peaks are a sign of disorder in many crystalline materials. Common sources of disorder—such as small grain size, defects or strain—produce well-known signatures in diffraction patterns. But none of these signatures seem to match the peak broadening exhibited by quasicrystals.

The X-ray diffraction results, in fact, point to a new form of structural disorder called phason disorder, which is unique to quasicrystals. If one compares a Penrose tiling or a quasiperiodic sequence with a conventional crystal, one sees that the quasiperiodic structures have the ability to generate a new



QUASIPERIODIC SEQUENCE of long and short segments (*top left*) is formed when a strip covers a lattice at a slope equal to an irrational number. Specifically, the slope is the inverse of the golden mean, or about 0.618. The sequence yields a sharp diffraction pattern (*top right*). A meandering path through the lattice produces a random sequence (*bottom left*). Despite this disorder, the diffraction pattern of the random sequence (*bottom right*) is similar to that of the quasiperiodic sequence.

form of disorder during growth: a defect can arise when the wrong kind of unit cell, or line segment, falls in a particular place. A few isolated mistakes will not affect the diffraction properties of an entire sample, but if many such mistakes plague a sample, they will disturb the diffraction patterns.

As an extreme example, imagine that the long and short segments in the one-dimensional quasiperiodic sequence are rearranged in a completely random fashion. Surprisingly, this random sequence gives rise to a diffraction pattern that is quite similar to the pattern derived from the original quasiperiodic sequence. The diffraction peaks of the random sequence are found at the same positions as those from the quasiperiodic sequence but are broader. In fact, the widths of these peaks are inversely related to the strength of the corresponding diffraction peak from the quasiperiodic sequence, so that only the more intense peaks remain. Nevertheless, the existence of relatively sharp diffraction peaks from the random sequence indicates that quasiperiodicity can survive disorder.

Such ideas led the authors to suggest in 1986 that the icosahedral alloys had an inherently defect-ridden structure. Our proposal became known as the icosahedral glass model. Randomness is important to the glass model in two ways. First, it removes the necessity of arcane matching rules and gives a more plausible explanation for quasicrystalline growth. Second, the disorder introduced through randomness closely mimics that evidenced by the peak broadening of the diffraction patterns. Interestingly, soon after the discovery of quasicrystals, Shechtman and Ilan Blech of Israel Institute of Technology–Technion in Haifa suggested that icosahedral quasicrystals were composed of icosahedral clusters that were randomly connected. This theory was refined by the authors to the point where we could reproduce the experimental observation of diffraction peak broadening [*see illustration on page 31*].

The attractiveness of the glass model extends beyond its ability to incorporate disorder. Certain crystalline alloys contain icosahedrally symmetric atomic clusters, which are plausible building blocks for the glass model. The size of those clusters is within 1 percent of that required to match the experimental diffraction patterns.

As these theoretical ideas were being developed in the late 1980s, materials scientists and chemists were busy in their laboratories discovering dozens of new icosahedral alloys. Some of the materials were variations of the aluminum-manganese alloys, but investigators also synthesized new families, such as aluminum-zinc-magnesium, uranium-palladium-silicon and nickel-titanium.

One of the most important results to come from these new materials was the discovery that quasicrystalline phases could be thermodynamically stable. The quenching process initially used by Shechtman and others produced very small grained quasicrystalline phases that, when heated, transformed irreversibly into common crystalline phases. Unfortunately, this metastability prevented workers from improving the quality of samples by heat treatment and other metallurgical techniques. Hence, the first quasicrystals had grain sizes of only a few thousandths of a millimeter, making many kinds of experiments impossible.

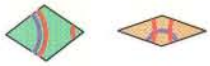
Several materials discovered in the past few years, however, retained their quasicrystalline structure up to their melting point. Hence, workers could prepare much larger samples by conventional crystal growth techniques. In this way, they have recently made alloys that have single grains as large as 10 millimeters in size.

Astonishingly, when these first stable quasicrystals of aluminum, lithium and copper were grown slowly enough to form large, faceted surfaces, they still suffered from the same degree of phason disorder as did their cousins formed by quenching. The discovery of phason disorder in these materials seemed to support the icosahedral glass model.

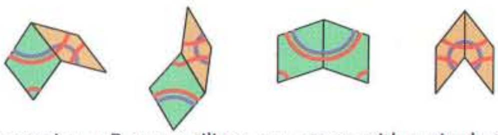
Although the icosahedral glass model is more successful at predicting the diffraction patterns, it goes too far in its attempt to incorporate disorder. It leaves too many gaps or tears in the structure where icosahedral clusters

How to Grow a Penrose Tiling

A Penrose tiling can be constructed from two kinds of rhombuses that have been decorated with colored bands.

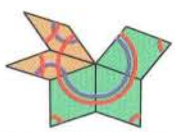


To make a Penrose tiling, one fits these rhombic tiles together according to the following "matching" rule: two rhombic tiles can be placed side to side only if bands of the same color join continuously across their interface. Some examples are shown below.

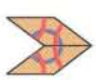


In growing a Penrose tiling, one starts with a single tile and adds each subsequent tile to the outer boundary. One assumes that once the tile has been positioned, it cannot be removed or shifted. (This growth process is somewhat similar to the formation of crystals, in which groups of atoms firmly attach themselves to the surface of a "seed" crystal.)

Growing a Penrose tiling is complicated by the fact that a given position on the boundary can often accommodate either of the two kinds of tiles. For example, a fat rhombus or two thin ones could be fit in the space at the top of these five tiles.



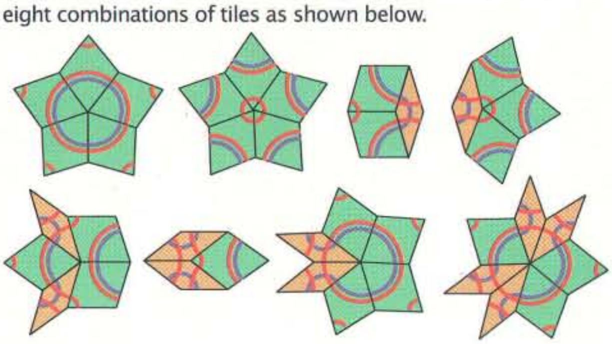
When these situations arise, one could choose among the possibilities at random. But such choices can lead to other problems. The rhombic tiles can be assembled so that they obey the matching rule but prevent other tiles from being added to the boundary. For example, neither a fat rhombus nor a thin rhombus can fit into the space at the left in the following configuration.



In general, if a wrong choice is made, the tiling will stop growing at some later point.

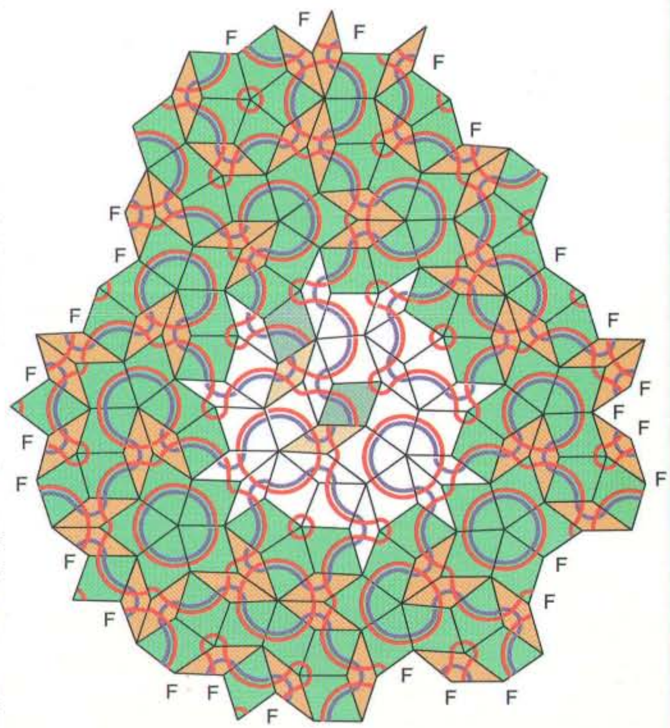
Can one develop a procedure that will allow Penrose tilings to grow indefinitely? Following the work of John H. Conway of the University of Cambridge and Nicolaas G. deBruijn of Eindhoven University of Technology, George Onoda of the IBM Thomas J. Watson Research Center and co-workers have taken the first step toward a solution.

To understand their method, one must first realize that each vertex in an infinite Penrose tiling is surrounded by one of eight combinations of tiles as shown below.

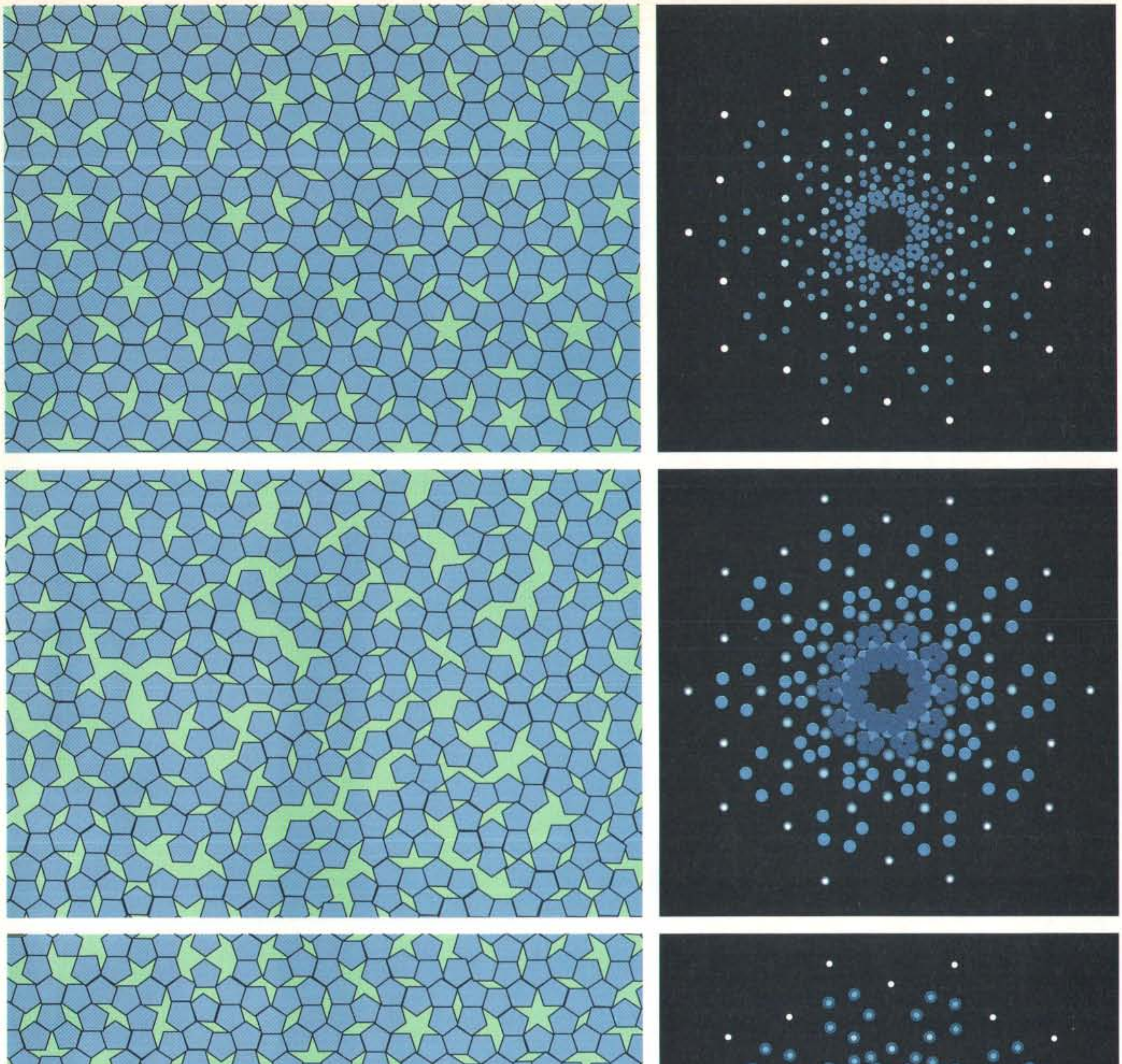


If a vertex on the boundary of a growing pattern will accommodate one or more tiles in only one way that is consistent with one of the eight vertices, then it is said that the vertex is forced. For example, the vertex indicated below by a dot is forced.

The boundary of a growing pattern will always contain at least one forced vertex if one starts with a particular seed in which the matching rules have been violated in certain places. Such a seed is represented by the white region below. The region contains two matching-rule violations that occur between the tiles shaded in light blue and yellow. As tiles were added to forced vertices on the boundary of the white region, the tiling began to grow, and new forced vertices were created on the new boundary. (Each forced vertex on the boundary in the pattern below is labeled with an *F*.) Because this growth process can be continued indefinitely, the tiling can be enlarged to any size. It will be a perfect Penrose tiling except for the original defects in the seed.



This procedure for growing Penrose tilings simulates only some aspects of real atomic growth. One can sensibly assume that forced vertices on the boundary represent sites where atoms attach easily to some surface. On the other hand, it seems unreasonable that some vertices will wait for an indefinite time before they become forced. It also seems odd that although several violations of the matching rules must occur in concert to produce an appropriate seed, additional mistakes must be avoided. Perhaps these issues will be resolved by further research.



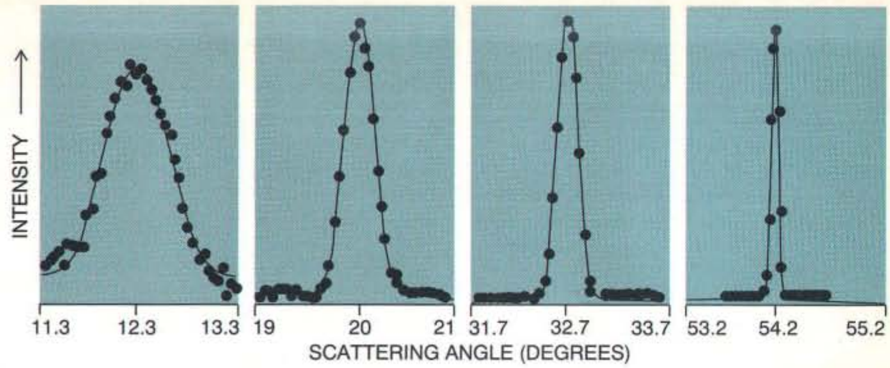
THREE MODELS have been proposed to account for the structure of quasicrystals and their diffraction patterns. The Penrose model is represented by the quasiperiodic lattice at the top left. The lattice is composed of pentagons, diamonds, stars and "boats" that have been assembled according to specific matching rules. The glass model yields the structure at the middle left. This structure is made of pentagons that have been stuck together, side by side, in a random way. The random-tiling model depicted at the bottom left is similar to the Penrose model, but it contains mistakes that are forbidden by the matching rules of the Penrose model. For each of these structures, one can calculate its diffraction pattern (*diagrams at right*), which can then be compared with experimental results. The Penrose and random-tiling model both yield sharp diffraction peaks, but the diffraction peaks from the random-tiling model (*bottom right*) are surrounded by faint halos.

cannot fit. These tears are, of course, absent in the Penrose model. The net effect of the tears is that the glass model overestimates the degree of broadening in diffraction patterns.

While the proponents of the Penrose and glass models debated the importance of local growth and phason disorder, a third group of investigators devised the random-tiling model, which combines some of the best concepts from the Penrose and the icosahedral glass models. The random-tiling model suggests that the strict matching rules of the Penrose model do not have to be obeyed, as long as there are no gaps left in the structure. Surprisingly, the random-tiling model predicts perfectly sharp diffraction peaks, just like its more ordered cousin, the Penrose model.

The apparent advantage of the random-tiling model is that it requires only local growth rules. For instance, Michael Widom, Katherine J. Strandburg and Robert H. Swendsen of Carnegie-Mellon University demonstrated that they could simulate the growth of these defect-filled tilings by applying the same computer algorithms used to simulate the growth of periodic crystals. Furthermore, they found that under certain circumstances the defect-filled tiling was more thermodynamically stable than a normal crystal. These investigators and Christopher L. Henley of Boston University demonstrated that the disorder associated with errors in a perfect quasiperiodic structure can actually stabilize quasicrystalline order, at least with respect to some competing crystalline phase. The relative importance of disorder increases with temperature, so that the random-tiling model predicts that quasicrystals achieve the stable, equilibrium phase only at elevated temperatures.

Over the past few years the three competing models for the icosahedral alloys have been refined to produce closer agreement with experiments and, in particular, diffraction data. For instance, theorists introduced mechanisms for producing disorder in ideal quasiperiodic structures to mimic more closely the broadened diffraction peaks. They developed algorithms to grow nearly perfect Penrose tilings by rules that seemed more plausibly local. At about the same time, Veit Elser, then at AT&T Bell Laboratories, modified the glass model by incorporating more realistic atomic motions during the simulated growth of a quasicrystalline grain; he found that the calculated diffraction patterns in such a grain did not exhibit excessive peak broadening beyond the experimental results. All these models were converging toward a middle ground incorporating some degree of phason disorder. To some observers, it seemed as though the differences between the models were dissolving into semantics.



DIFFRACTION PROPERTIES of a quasicrystal composed of aluminum, copper and lithium were revealed by bombarding the quasicrystal with a beam of X rays and by measuring the intensity of the scattered X rays at various angles relative to the beam. The graph shows that the diffraction peaks sharpened with increasing angle. This sharpening is the signature of phason disorder, a kind of structural defect found only in quasicrystals. The measurements were taken by Paul Heiney, Paul Horn, Frank Gayle and their co-workers at the National Synchrotron Light Source.

A series of experiments performed in 1989, however, essentially narrowed the field of plausible models to two: the Penrose and random-tiling models. Physicists at Tohoku University in Japan discovered a new family of icosahedral alloys, including aluminum-copper-iron and aluminum-copper-ruthenium. Remarkably, when we took X-ray diffraction patterns of these materials, we found that the peak broadening associated with phason disorder—an effect evident in all previously studied icosahedral alloys—was absent. Peter Bancel of the IBM Thomas J. Watson Research Center independently confirmed these results and showed that as the iron alloy was heated and cooled, phason peak broadening could be enhanced or diminished.

Is quasicrystalline order produced by the matching rules of the Penrose model or by the freedom to scramble local groups of atoms to allow limited phason disorder? In an attempt to answer this question, several laboratories, including our own, are continuing experiments to study the conditions that control the perfection of quasicrystals.

While some investigators have been working to fathom the novel forms of atomic order displayed by these materials, others have been developing applications. Yi He, S. Joseph Poon and Gary J. Shiflet of the University of Virginia have used insights about quasicrystalline structure to synthesize metallic glasses containing up to 90 percent aluminum. They hope to exploit the low density and unusually high strength of these materials. Jean-Marie Dubois and his colleagues at the National School of Mining Engineering in Nancy, France, have discovered that certain quasicrystals produce excellent low-friction coatings.

Quasicrystals continue to pose exciting challenges for condensed matter scientists. Nearly all the ideas that have been developed to understand the electronic, thermal and mechanical properties of crystalline solids are based on the simplifying framework of periodicity. Now faced with quasiperiodic structures, we must seek more sophisticated levels of understanding.

FURTHER READING

ICOSAHEDRAL CRYSTALS: WHERE ARE THE ATOMS? Per Bak in *Physical Review Letters*, Vol. 56, No. 8, pages 861–864; February 24, 1986.

QUASICRYSTAL EQUILIBRIUM STATE. Michael Widom, Katherine J. Strandburg and Robert H. Swendsen in *Physical Review Letters*, Vol. 58, No. 7, pages 706–709; February 16, 1987.

INTRODUCTION TO QUASICRYSTALS. Edited by Marko V. Jarić. Academic Press, 1988.

Al-Cu-Ru: AN ICOSAHEDRAL ALLOY WITHOUT PHASON DISORDER. C. A. Guryan, A. I. Goldman, P. W. Stephens, K. Hiraga, A. P. Tsai, A. Inoue and T. Masumoto in *Physical Review Letters*, Vol. 62, No. 20, pages 2409–2412; May 15, 1989.

DYNAMICAL PHASONS IN A PERFECT QUASICRYSTAL. Peter A. Bancel in *Physical Review Letters*, Vol. 63, No. 25, pages 2741–2744; December 18, 1989.

EXTENDED ICOSAHEDRAL STRUCTURES. Edited by Marko V. Jarić and Dennis Gratias. Academic Press, 1989.

Molecular Zippers in Gene Regulation

Recurring copies of the amino acid leucine in proteins can serve as teeth that "zip" two protein molecules together. Such zippering plays a role in turning genes on and off

by Steven Lanier McKnight

Like contemporary scientists, the cells of humans and other multicellular organisms tend to specialize. Consider, for example, skin cells and liver cells. The skin cells known as keratinocytes form a protective barrier between an animal and its surroundings. They are quite different from the hepatocytes of the liver, which store glycogen (a source of energy), eliminate toxins from the blood and secrete many constituents of blood serum, including albumin. The question of how such differences arise has challenged biologists for the better part of a century.

A somewhat superficial answer might be that the identifying properties of cells derive from the specialized proteins they produce, as distinct from the "housekeeping" proteins made in every cell. For instance, keratinocytes gain their strength from fibrous proteins called keratins, which assemble into long, intertwined cables. Similarly, hepatocytes can store glycogen because they synthesize liver-specific enzymes tailored to that task.

At a deeper level, one might add that because the instructions for making proteins are carried in the genes, the defining properties of a cell are determined by the mix of genes that are active. (Each gene, which is made up of DNA, specifies a single protein.)

STEVEN LANIER MCKNIGHT is a staff member in the department of embryology of the Carnegie Institution of Washington in Baltimore and is an investigator at the facility's Howard Hughes Medical Institute. He earned his Ph.D. in biology from the University of Virginia in 1977. McKnight has been affiliated with the Carnegie Institution ever since, except for the years 1981 to 1984, which he spent at the Fred Hutchinson Cancer Research Center in Seattle.

Even this more precise answer begs the question, because it does not explain why certain genes are activated in one cell but not in another. Virtually all cells in an organism carry identical genes. Yet only keratinocytes abundantly express keratin genes, that is, they transcribe the genes into messenger RNA templates for the keratins and then translate those templates into the keratins themselves. Likewise, only hepatocytes express the albumin gene.

Thus, the problem of differential gene expression continues to challenge, but it is slowly yielding to scrutiny, as is the broader question of how any gene is switched on and off. In our own contribution to such work, my colleagues and I at the Carnegie Institution of Washington have built on the major discovery, made by other workers, that gene expression is controlled to a great extent by proteins that bind to DNA.

We have demonstrated that many such gene-regulating proteins "zip" together into pairs. This linkage is critical to their ability to bind to DNA. It also seems to control gene activity in other ways and to play a part in determining why a gene is turned on in one cell but not in another. Because the "teeth" that join the molecules almost always consist of the amino acid leucine, we have come to call the toothed region the leucine zipper.

We discovered the leucine zipper in 1987, as we were trying to learn more about how DNA-binding proteins activate genes. The first such proteins had only recently been identified. Investigators knew something about how they worked, but an impasse of sorts had been reached. To understand that impasse, it helps to have a sense of the state of research at the time.

Gene regulatory proteins were discovered as an outgrowth of research into the structure of DNA and, later, into the organization of genes. By the early 1980s scientists had known for more than 25 years that DNA is a double helix, composed of two strands of nucleotides (the building blocks of DNA). Nucleotides consist of a sugar, a phosphate and chemical group called a base—either adenine (*A*), cytosine (*C*), guanine (*G*) or thymine (*T*). The two DNA strands are linked at the bases, and the nucleotides thus coupled are known as base pairs. Adenine on one strand always pairs with thymine on the complementary strand, and cytosine pairs with guanine.

Investigators had also determined that genes include both a protein-encoding element and regulatory elements. The base pairs of the encoding element specify the amino acids to be linked together into a protein chain—the gene product. The regulatory elements control transcription of the encoding segment. Work in bacteria, and subsequently in multicellular organisms, established that one regulatory element, called the promoter, lies adjacent to the encoding region. The promoter dictates where on the DNA molecule an enzyme called RNA polymerase will initiate transcription, and it influences the rate of transcription.

During the 1980s, studies of viruses uncovered other regulatory elements called enhancers. An enhancer can lie thousands of base pairs from the encoding element it controls. In other respects, though, enhancers seem to be quite similar to promoters [see "DNA," by Gary Felsenfeld; SCIENTIFIC AMERICAN, October 1985]. For example, enhancers, like promoters, increase the rate of transcription. Indeed, they work in concert with promoters. Both elements can consist of several hundred base pairs, which are grouped into mo-

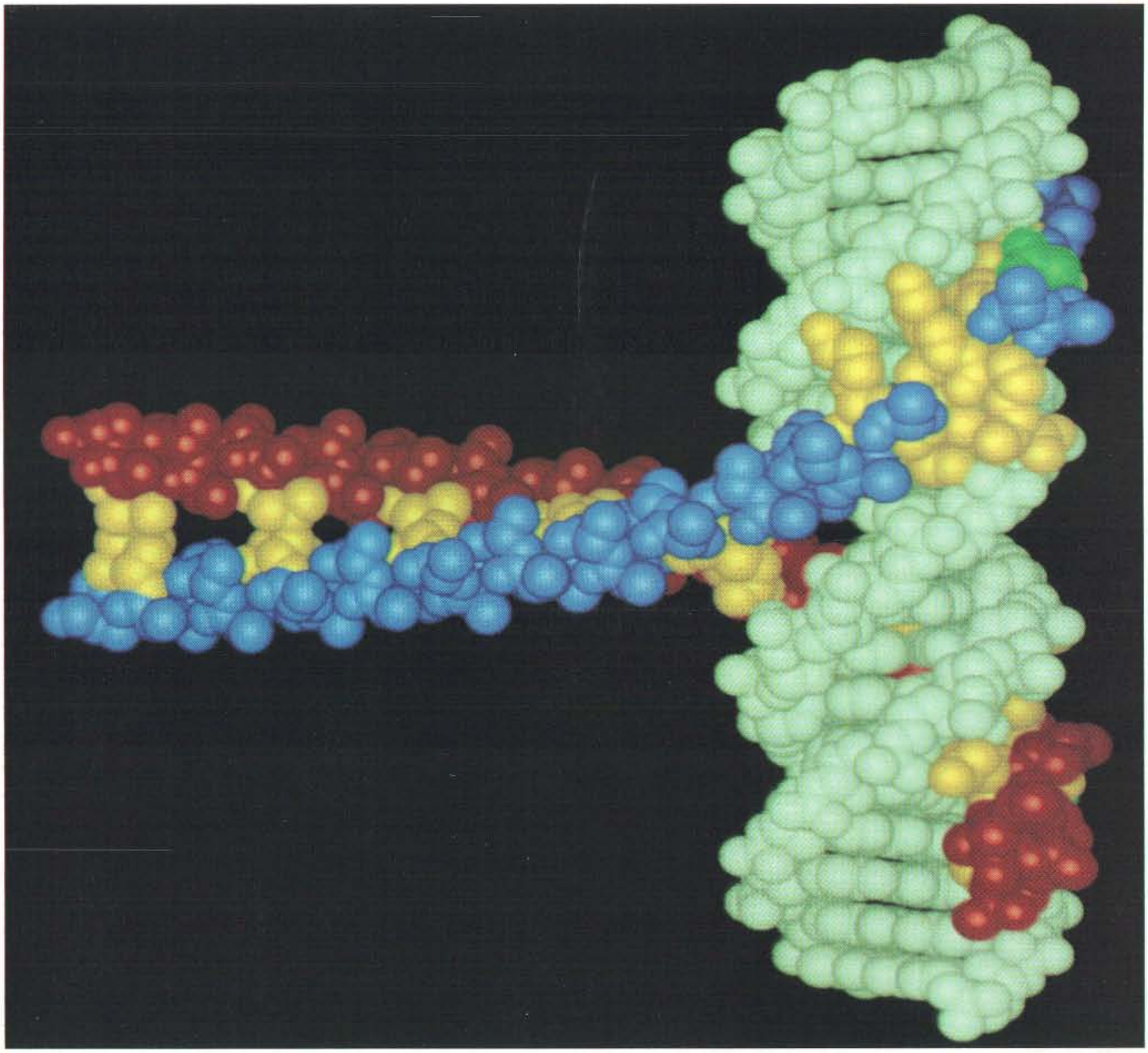
tifs: distinctive sequences of six to 10 base pairs. A typical enhancer or promoter might include five to 10 motifs, some of which repeat.

I should note that other regulatory elements known as silencers, which repress gene expression, have also been discovered. Elegant work by Alexander D. Johnson of the University of California at San Francisco has shown that silencers, like enhancers, can function at great distances from the encoding element of a gene and consist of a patchwork of DNA motifs. They are also probably controlled in much the same way as enhancers. Although silencers are important to gene regulation, I shall concentrate in this article on the elements controlling the activation of genes.

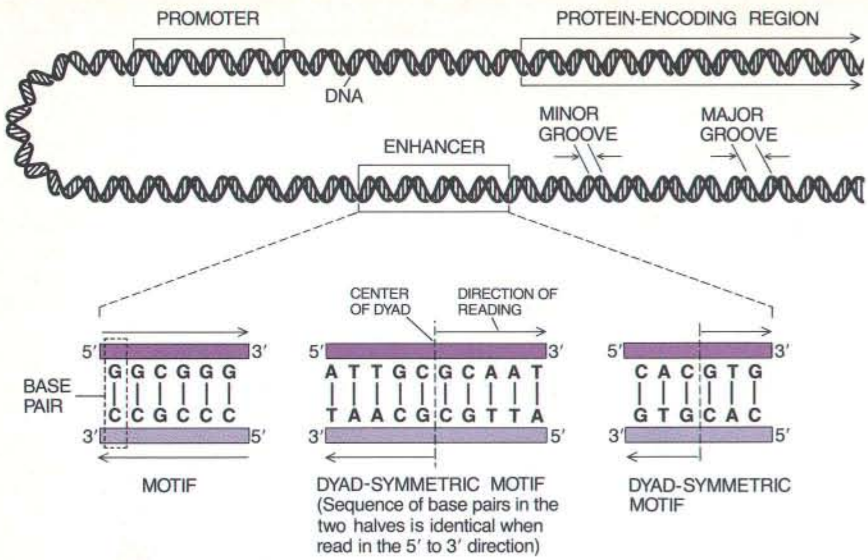
When promoters and enhancers were first identified, little was known about how they worked. Then, in about 1984, Robert Tjian and his colleagues at the University of California at Berkeley made a fundamental breakthrough. They found that a motif repeated five times in the promoter of what is called the *early* gene in the mammalian SV40 virus could be avidly bound by the protein SP1. Moreover, after SP1 bound it, transcription of the *early* gene was selectively activated. This was the first demonstration outside of bacteria that regulatory elements on genes could be controlled by proteins bound to them. The list of such regulatory proteins, which includes many in multicellular organisms, continues to grow.

Tjian's discovery gave rise to a plausible model of gene activation. In that view, each motif in a promoter or enhancer represented a binding site for a regulatory protein. A gene would be expressed significantly only if every motif in its promoter and enhancer was bound. Hence, expression of a particular gene would be restricted to cells capable of synthesizing the complete array of regulatory proteins needed to recognize every motif.

The prediction that individual motifs



ZIPPERING of two proteins (*left part of image*) by interaction of their leucine amino acids (*yellow spheres*) can help certain proteins to grip DNA (*white*) and thus to activate or silence genes. The yellow spheres touching the DNA (*right*) represent not leucine but positively charged amino acids that may strengthen the proteins' grip. The green cluster is the amino acid asparagine; it may help the DNA-binding segments, which track in the wide major groove, to wrap around DNA.



TYPICAL GENE includes a protein-encoding region and two elements, called the promoter and enhancer, that regulate gene transcription: the copying of the encoding region into messenger RNA for translation into a specified protein. Promoters and enhancers consist of several motifs, groupings of nucleotides (the building blocks of DNA) that constitute docking sites for proteins. The nucleotides themselves are distinguished by chemicals called bases—adenine (*A*), cytosine (*C*), guanine (*G*) or thymine (*T*). Zippered proteins often bind to dyad-symmetric motifs, which contain two identical halves; each protein in a pair binds to one of the two halves.

are binding sites for regulatory proteins turned out to be correct, as did the notion that every motif has to be occupied if a gene is to be regulated appropriately. Yet, in a surprising turn of events, my colleagues and I noticed that although only hepatocytes synthesize albumin, cells taken from the brain and spleen also contained proteins that could recognize the known regulatory elements of the albumin gene.

This and similar observations by others suggested that the steps involved in switching on a gene were more complex than originally envisioned. Protein recognition of every regulatory motif in a gene apparently was not adequate in itself. To gain a fuller understanding of how promoters and enhancers worked, investigators would have to learn much more about the proteins that bind DNA.

My colleagues and I were among the many scientists who took on that task. We decided to concentrate on a protein that recognized a motif called CAT (short for CCAAT) found in many promoters, both in viral and mammalian genes. Once we isolated a small amount of the protein, we planned to raise an antibody against it. The antibody would serve as a tracer, enabling us to pinpoint when and where in the mammalian body the protein was put to work.

By more circuitous means, the antibody would also help us to clone the gene encoding the protein—that is, to isolate the gene and synthesize its product in quantity. That accomplished, we would be able to study the regulatory protein in detail, fishing around for clues as to how it helps to activate other genes. We would be able to determine its complete amino acid sequence and perhaps deduce something about its three-dimensional structure, the mechanics of its binding to DNA and the nature of its interactions with other regulatory proteins.

The plan was sound but difficult to carry out. Indeed, my colleagues Peter F. Johnson, Barbara J. Graves, William H. Landschulz and I spent three years just purifying the protein and raising an antibody against a small fragment of it.

As soon as we had purified the protein, we found that it had affinity not only for the CAT motif in promoters but also for a motif called the core homology, common to many enhancers. We thus named the substance C/EBP, for CAT/Enhancer-Binding Protein.

With the help of the antibody, we also discovered that our protein was not made in all tissues. It was abundant in the liver, lungs, small intestine and placental tissue as well as in fat. But it was virtually absent in most other tissues of adult mammals. Where it did occur, it was restricted to the specialized cells that define the physiological properties of tissues. For example, C/EBP was made by hepatocytes but not by cells that form the secretory duct work of the liver.

This last observation raised the possibility that C/EBP might participate in selective gene expression, helping to switch on the genes for the specialized proteins (such as albumin) that make one cell distinct from another. We soon confirmed that C/EBP did contribute to the production of specialized proteins in hepatocytes and fat cells. We had managed to isolate just the kind of regulatory protein we would need if we were to learn how differential gene expression is regulated in mammals.

As soon as we had cloned the C/EBP gene and determined its full nucleotide sequence, thus revealing the identity and order of the protein's 359 amino acids, we immediately ran the amino acid sequence through a computer data base to see whether it resembled that of any other known proteins. The finding of similar sequences in parts of other proteins would guide us to regions that were potentially important to the function of C/EBP. When specific stretches of amino acids have crucial roles, they are often conserved during evolution, so that the same stretch reappears in functionally related molecules within and also between species.

The computer search revealed that a 60-amino acid segment of C/EBP was quite similar to segments of two other proteins: the products of the *myc* and *fos* proto-oncogenes. Proto-oncogenes are genes that normally serve the body well but become cancer-causing when they undergo certain mutations. In C/EBP, some part of the highlighted region was known to participate in binding DNA. The normal activities of the Myc and Fos proteins were not known, but the fact that a part of them resembled the C/EBP DNA-binding region suggested they too might be gene regulatory proteins. Subsequently, other workers showed that Myc and Fos are indeed regulators.

We, in the meantime, set about deciphering the architecture of the related regions. The three-dimensional structure of a protein, which is dictated by its sequence of amino acids, influences how it interacts with other molecules. Unfortunately, there are no foolproof rules for predicting a protein's architecture from its linear amino acid sequence, and so we had to resort to a certain amount of guesswork [see "The Protein Folding Problem," by Frederic M. Richards; SCIENTIFIC AMERICAN, January].

We knew that parts of many proteins fold into an alpha helix, a kind of coil. We therefore began by considering whether the segments relating C/EBP, Fos and Myc might adopt that structure. If the regions contained the amino acids proline or glycine, an alpha helix would be unlikely, because those two amino acids are rarely found in the helices. In fact, there were no prolines or glycines.

Amino acids with different chemical properties are often segregated on the surface of alpha helices, a feature known as amphipathy. For example, a helix might have hydrophobic (water-hating) amino acids on one face and hydrophilic (water-loving) versions on the opposite face. Evidence of such amphipathy would support our proposal of helicity, and so Landschulz examined a 35-amino acid segment of C/EBP to see if it would be amphipathic when arranged as an alpha helix. Sure enough, the hydrophobic and hydrophilic amino acids were segregated.

Because hydrophobic amino acids are attracted to other water-hating substances, it seemed possible that the hydrophobic face might interact with hydrophobic regions on other proteins, whereas the hydrophilic face would "be content" to interact with water, the major constituent of cells. Ultimately we confirmed that the hydrophobic face does interact with other proteins, but in an unusual way.

I remember scanning a penciled sketch of the putative C/EBP helix that Landschulz had completed just before the Christmas holiday of 1987. I was surprised to see that the amino acid leucine, which is extremely hydrophobic and protrudes relatively far from the backbone of the helix, occupied every seventh position. This "heptad repeat" caused the leucines to be aligned in a plane along the length of the helix, so that they essentially formed a continuous ridge.

The orderly arrangement of the leucines intimated that they served some useful purpose; perhaps they participated in the hydrophobic interactions we had posited. Landschulz and I then inspected the Myc and Fos proteins to see if they too displayed the same pattern, which would bolster the possibility that the heptad repeats were important. To our delight, we saw that the C/EBP-related regions of Myc and Fos not only were amphipathic when fit into an alpha helix, they also sported the same heptad repeat of leucines seen in C/EBP.

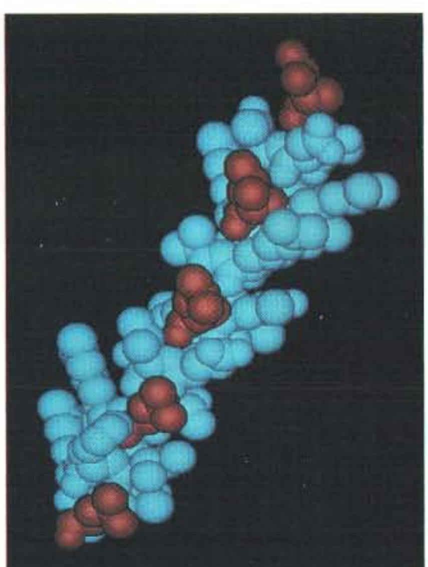
We now know that the leucine repeats in gene regulatory proteins enable such proteins to join at their zipper regions, forming two-part units called dimers. That understanding took time to gel, however, and depended on certain other discoveries.

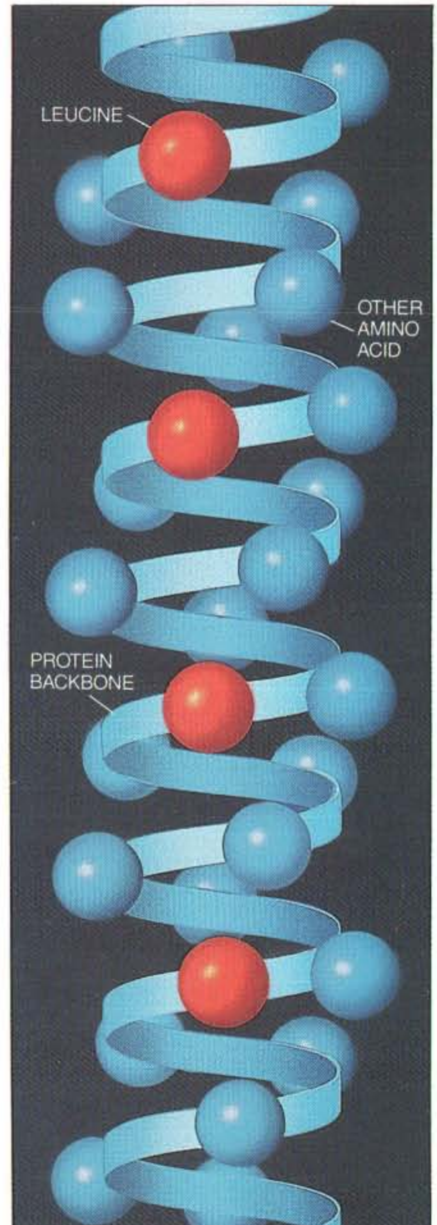
First, in an observation that was to influence our thinking critically, we noticed that the motifs bound by C/EBP are dyad symmetric, a feature common to many promoters and enhancers. Such motifs consist of two identical halves [*see illustration on opposite page*]. Mark Ptashne and his colleagues at Harvard University had already demonstrated in bacteria that such motifs are bound by dimers and that each dimer subunit interacts with one of the dyad halves [see "A Genetic Switch in a Bacterial Virus," by Mark Ptashne, Alexander D. Johnson and Carl O. Pabo; SCIENTIFIC AMERICAN, November 1982]. Their work suggested to us that C/EBP might combine with itself to form a dimer before binding to DNA.

Second, we learned of two additional proteins that were related to C/EBP. One of them, called GCN4, was a gene regulatory protein from yeast. In what was becoming a recurring theme, Kevin Struhl and his colleagues at Harvard had determined that the DNA motif bound by GCN4 was dyad symmetric and that GCN4 hooks onto the DNA as a dimer.

The other protein, which was identified by Peter K. Vogt and his co-workers at the University of Southern California, was the product of a proto-oncogene known as *jun*. Provocatively, Tom Curran of the Roche Institute of Molecular Biology had shown that the Jun protein could combine with the Fos protein to form a dimer. A knowledge of how the combination took place would presumably help clarify how the two proteins function in the cell.

The C/EBP-related regions of both GCN4 and Jun contained leucine at every seventh position of an idealized alpha helix. Putting all of these observations together, Landschulz, Johnson and I proposed that the role of the amphipathic alpha helix was to create a



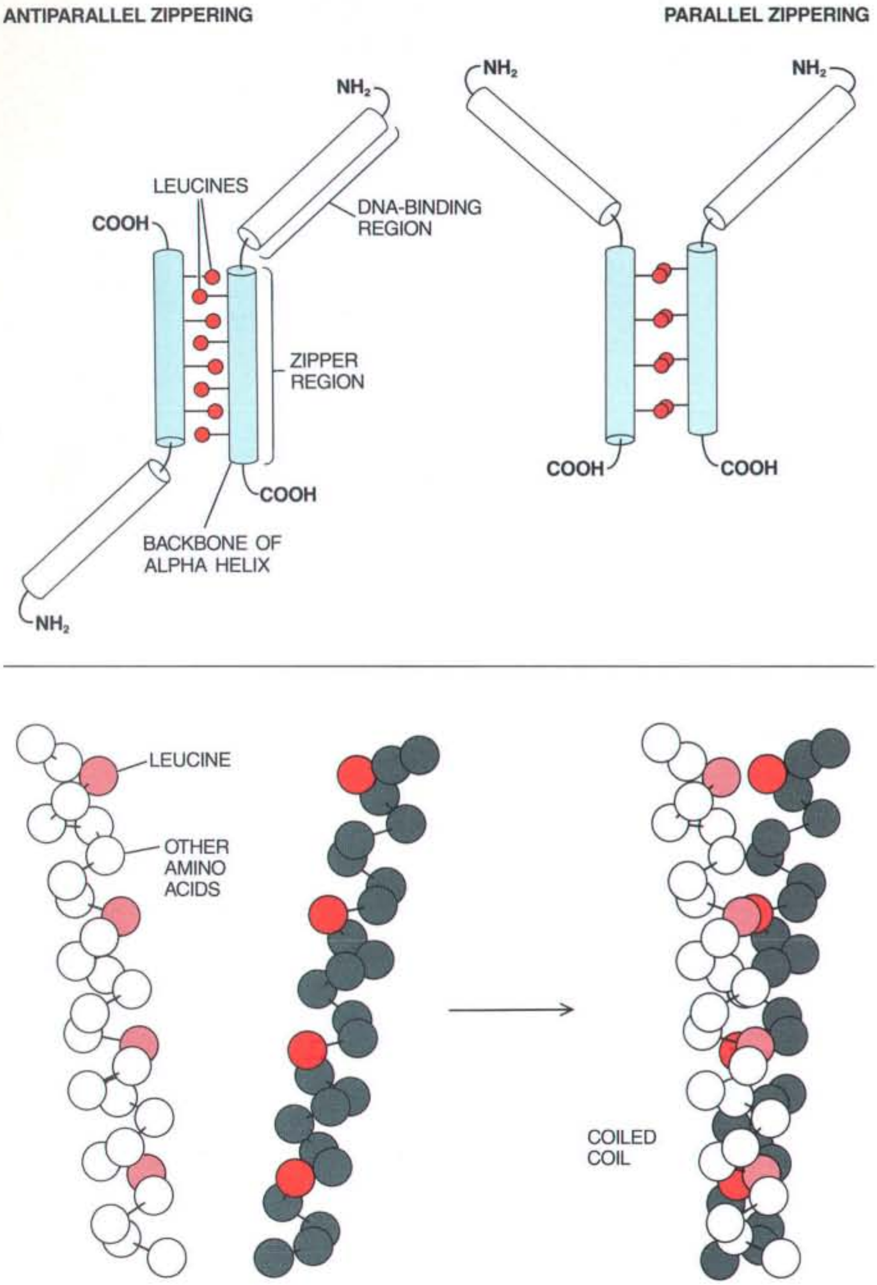


LEUCINE ZIPPER (*above*) was discovered when a small segment of the protein C/EBP was fit into a hypothetical alpha helix, a structure in many proteins. Surprisingly, the leucines, which make up every seventh amino acid, lined up in a column. A computer model (*left*) shows another view of the putative helix. The red projections are leucines; the blue spheres represent other amino acids.

surface that would enable proteins to coalesce into dimers and that such dimer formation was required for DNA binding. Because leucines strongly attract one another, we further postulated that the ridge of leucines on one protein would interlock with a similar ridge on another protein, essentially zippering the two proteins together.

Many findings now support these proposals. Notably, mutations that would disrupt the creation of an alpha helix in the zipper section prevent zippering. Dimer formation is also blocked by mutations that exchange leucines with amino acids that generate weaker hydrophobic attractions. Importantly, all mutations that impair dimerization also block DNA binding.

Initially we also suspected that the zipper regions might adopt an antiparallel orientation when they joined, so that the leucines would interlock like the teeth of a true zipper. Since then, however, Peter S. Kim and Erin K. O'Shea of the Whitehead Institute for Biomedical Research at the Massachusetts Institute of Technology, applying a variety of investigative techniques to the GCN4 molecule, have established that zipper regions associate in parallel when they combine. Instead of interdigitating, the leucines on apposed molecules line up side by side [*see illustrations on this page*].



HOW TWO PROTEINS MESH at their zipper regions was not initially obvious. If the regions were antiparallel, the leucines would be expected to interdigitate (*top left*) like the teeth in a true zipper. If the molecules were parallel, the opposing leucines would instead overlap (*top right*). The parallel model, which places the DNA-binding regions in a suitable position for contacting dyad-symmetric motifs, is now known to be correct. The proteins combine to form what is called a coiled coil (*bottom*).

Kim and O'Shea further showed that the joined helices form a classic coiled coil—a well-known structure seen in dimers of many long fibrous proteins, including keratins and the lamins that form the nuclear envelope. The fibrous proteins zipper by means of regularly spaced hydrophobic amino acids yet are not dependent solely on leucine, the quintessential hydrophobic amino acid. Hence, the leucine zipper is a variation on a widespread theme; it probably evolved to ensure the tight linkage of short alpha helices.

In yet another prediction, we suggested that zippers could either join two identical proteins, forming what is called a homodimer, or join dissimilar proteins, such as Fos and Jun, to form a heterodimer. The idea that two different proteins could be coupled by the zipper was the most unconventional aspect of our model, but various investigators have now proved it correct. Included in the known mixed pairs are heterodimers formed by the combination of different variants of the C/EBP protein.

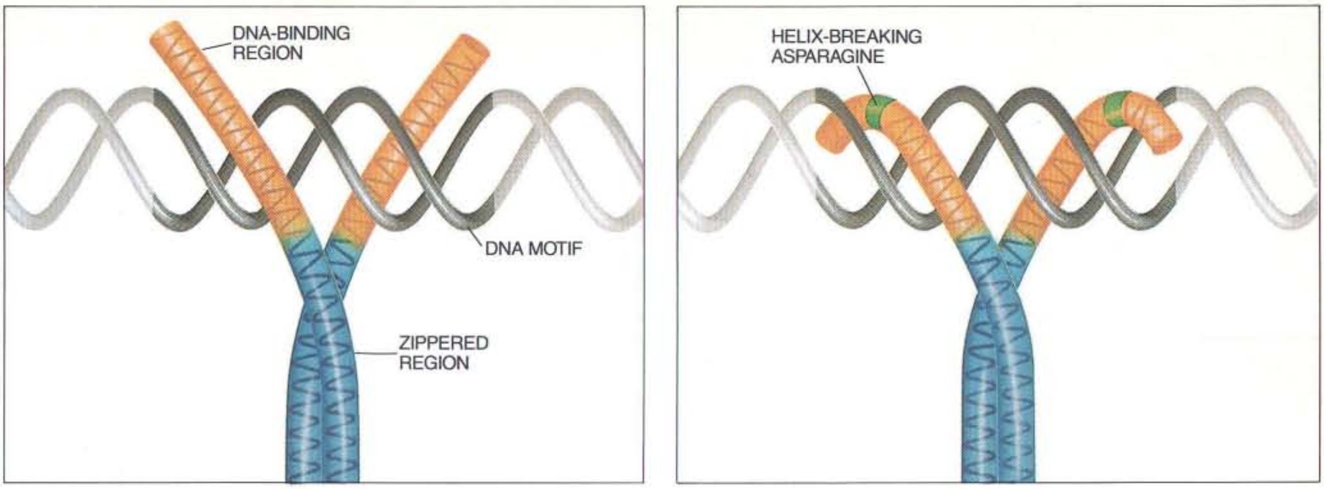
We recognized that such cross-mixing could be important to the ability of an organism to regulate gene expression. To better understand the specific role of the cross-mixing, though, we first needed to know more about how zippering might aid binding. Although the leucine zipper can generate the dimers needed for DNA binding, the zipper region cannot itself bind DNA.

Early in our studies of C/EBP we realized that a region rich in basic (positively charged) amino acids—specifically, arginine and lysine—resided close to the zipper. Because DNA is quite acidic (negatively charged) and because positively charged substances strongly attract negatively charged ones, we postulated that this basic, or arg/lys, region might be the part of the protein that directly contacted DNA. If that was the case, perhaps the zippering of two molecules served to bring their arg/lys regions into proper position for combining with dyad-symmetric motifs.

Studies have now shown that the arg/lys region is indeed the DNA-binding region. For instance, mutations in that region have been found to prevent

EARLY MODEL OF DNA BINDING

CURRENT MODEL



SOURCE	NAME OF GENE REGULATOR	AMINO ACID SEQUENCE (each letter represents an amino acid): DNA-BINDING REGION / 6–AMINO ACID CONNECTOR / LEUCINE ZIPPER
MAMMAL	C/EBP	DKNSNEYRVRRERNNIAVRKSRDKAKQRNVETQQKVLELTSDNDRLRKRVEQLSRELDTLRG–
	CREB	EEAARKREVRLMKNREAARECRRKKKEYVKCLENRVAVLENQNKTLIEELKALKDLYCHKSD–
	JUN	SQERIKAERKRMRNRIAASKCRKRKLERIARLEEKVKTLKAQNSELASTANMLTEQVAQLKQ–
	FOS	EERRRIRRIRRERNKMAAAKCRNRRRELTDTLQAETDQLEDKKSALQTEIANLLKEKEKLEF–
YEAST	GCN4	PESSDPAALKRARNTEAARRSRARKLQRMKQLEDKVEELLSKNYHLENEVARLKKLVGER–COOH
	YAP1	DLDPETKQKRTAQNRAAQRAFRERKERKMKELEKKVQSLESIQQQNEVEATFLRDQLITLVN–
OTHER FUNGI	CYS-3	ASRLAAEEDKRKRNTAASARFRIKKKQREQALEKSAKEMSEKVTQLEGRIQALETENKYLKG–
	CPC1	EDPSDVVAMKRARNTLAARKSRERKAQRLEELEAKIEELIAERDRYKNLALAHGASTE–COOH
PLANT	HBP1	WDERELKKQKRLSNRESARRSRLRKQAECEELGQRAEALKSENSSLRIELDRIKKEYEELLS–
	TGA1	SKPVEKVLRRLAQNREAARKSRLRKKAYVQQLENSKLKLIQLEQELERARKQGMCVGGVDA–
	OPAQUE 2	MPTEERVRKRKESNRESARRSRYRKAAHLKELEDQVAQLKAENSCLLRRIAALNQKYNDANV–
	CONSENSUS MOLECULE	---------RR/KK-R/K-N------R/K--R--RR/KK------L------L------L------L------L---

ARGININE

LYSINE

INVARIANT ASPARAGINE

LEUCINE

DNA-BINDING REGION of zipper proteins was once thought to be a continuous alpha helix and to protrude from the DNA (*top left*). A comparison of the amino acid sequences of 11 proteins (*bottom*) showed that certain amino acids in the protrusion were conserved from molecule to molecule (*yellow highlighting*), suggesting that they aid binding and touch the DNA. Moreover, asparagine (*green*), a potential helix breaker, was found to lie at a fixed point in the DNA-binding region of every protein. It, then, might allow the protruding section to bend and thus contact the DNA continuously (*top right*).

dimers from combining with DNA efficiently. Research into the structure of the binding region has also supported the notion that zippering orients the arg/lys regions appropriately for contact with DNA.

Much of our knowledge of how zippered proteins bind DNA came from an extensive comparison of 11 gene regulatory proteins that each contained neighboring zipper and arg/lys regions. The selection included proteins from plants, fungi and mammals—organisms that in some cases (notably plants and mammals) diverged more than a billion years ago.

The comparison initially suggested that all the proteins were designed so that when they were assembled into a dimer, the resulting unit would adopt the shape of a Y. The zippered region would form the stem, of course, and the arg/lys regions would form the DNA-binding arms.

We further suspected that the helix of the zipper region would extend straight through the DNA-binding region. We thought so in part because the arg/lys segments of all 11 proteins were free of the helix-disrupting amino acids glycine and proline. Moreover, we saw that in every one of the proteins, the zipper and DNA-binding region were separated by exactly six amino acids. If that connector segment was helical, it would serve to ensure that any conserved amino acid in the arg/lys area of the helix would lie on the same face in every molecule—presumably on the face involved in DNA binding.

We anticipated as well that each helical arm of the Y would combine with

one half of a dyad-symmetric recognition site on DNA. The DNA helix itself is grooved, much as a screw has grooves. The major groove is wide, and the minor groove is narrow. We expected that one arm of the Y would fit into the major groove on the near face of the DNA, touching one half of the motif, and the other arm would fit into the major groove of the far face, touching the other half of the motif.

To evaluate the validity of this conceptual model, Charles R. Vinson, a postdoctoral student in my laboratory, built a ball-and-stick model pairing two identical molecules of C/EBP. The molecules zippered nicely and could be made to adopt the Y shape.

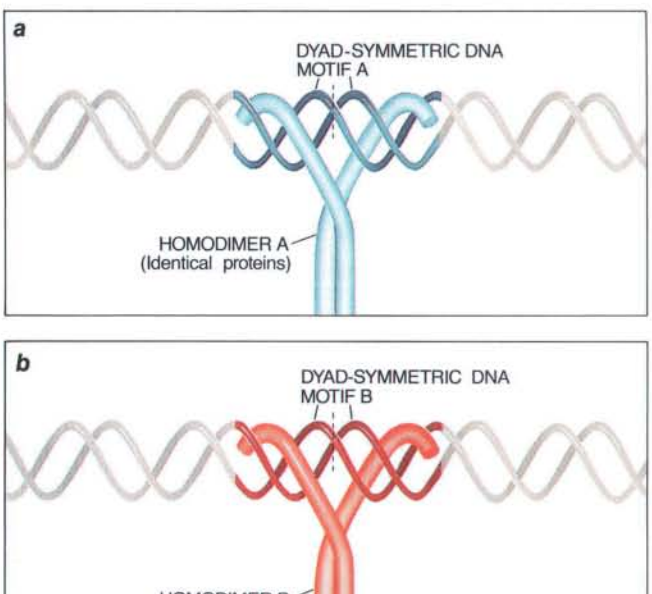
In Vinson's model the conserved arginines and lysines fell on the inside of each arm of the Y. That arrangement fit well with our conception, because it placed the arginines and lysines in a suitable position for forming ionic bonds with the negatively charged sugar and phosphate backbone of the DNA. Such bonding would help keep the protein firmly in place on the DNA.

A second finding was less supportive of our initial thinking. We saw that the arg/lys region, if displayed as a continuous helix, would be too long to allow uninterrupted contact with DNA. The top of each arm of the Y would protrude from the DNA instead of folding around it. Such an arrangement seemed unlikely considering that the arginines and lysines in the protruding segment are highly conserved and therefore almost certainly involved in binding DNA. Moreover, we found that substitution of those amino acids with others prevented binding.

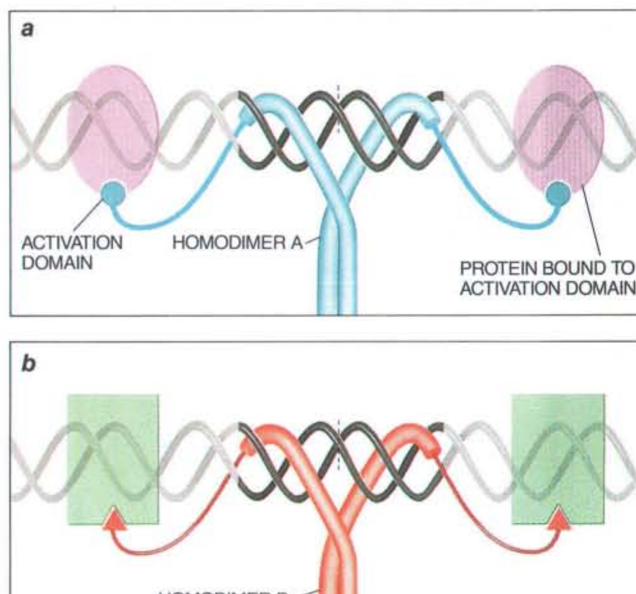
Thus, we were forced to acknowledge that if the arg/lys region was in fact helical, the helix would have to be broken at some point, enabling it to bend. Puzzled, we reinspected our 11 proteins. Although the arg/lys regions lacked the classical helix breakers glycine and proline, they all contained the amino acid asparagine in the same spot. Like proline and glycine, asparagine can interrupt helices.

With the help of Paul B. Sigler and his colleagues at Yale University, we refined our molecular modeling to include a break in the helix at the invariant asparagine. The new model enabled the full length of the arg/lys regions, including the formerly protruding parts, to track continuously in the major groove of the DNA molecule. The binding of DNA by those regions of a protein dimer is analogous to the scissors grip applied by wrestlers as they lock their legs around the torso of opponents.

HETERODIMERS INCREASE NUMBER OF USABLE MOTIFS



HETERODIMERS INCREASE VARIETY OF PROTEIN COMBINATIONS



HETERODIMERS, zippered pairs of nonidentical proteins, offer two possible benefits to an organism. If homodimers, pairs of identical proteins, recognize only dyad-symmetric (twinlike) motifs (*a* and *b at left*), the existence of heterodimers able to recognize asymmetric (mismatched) motifs (*c*) would increase the number of motifs that can be enlisted for gene regulation. Alternatively, the benefit could lie elsewhere (*right*). Protein-binding "activation domains" in a homodimer always combine with identical proteins (*a* and *b*), whereas the domains of heterodimers bind dissimilar proteins (*c*). The heterodimers may thus increase the number of ordered protein complexes that can be used to turn genes on or off.

Having established that many proteins zipper to form dimers, my collaborators and I began to speculate on why evolution might have built the zipper feature into many gene regulators. In our view, the chief value of this feature may well lie with the creation of heterodimers by the cross-mixing of nonidentical proteins.

To explain the logic of this view, I must update the model of gene regulation discussed earlier. Recall that if a gene is to be activated, every DNA motif in its promoter and enhancer has to be bound by a regulatory protein. Yet it seems that such attachment alone is not sufficient. Apparently, DNA-binding proteins must also "fit" comfortably with proteins that bind to neighboring motifs.

Although several different proteins seem capable of recognizing and attaching to a particular motif, so that several different combinations of molecules can in theory cover a particular enhancer or promoter, many researchers now suspect that only one of the possible permutations will in fact activate the promoter or enhancer of a given gene.

In other words, only one combination of regulatory proteins will mesh into a "key" having the correct three-dimensional shape to regulate transcription. Cells that lack even one piece of the key, which I sometimes liken to a jigsaw puzzle, will fail to activate the corresponding gene. Hence, a gene will remain silent if a cell is missing, say, just one subunit of a crucial dimer or a protein that interacts appropriately with that subunit.

According to this thinking, the series of motifs in an enhancer or promoter form the template to the puzzle. The actual pieces of the puzzle are the gene regulatory proteins themselves. Exactly how such jigsaw puzzles assembled on DNA enable transcription to occur is not clear but is the focus of intense study [see "How Gene Activators Work," by Mark Ptashne; SCIENTIFIC AMERICAN, January 1989].

Within this conceptual framework, I can think of two benefits that heterodimeric regulatory proteins might offer an organism. In one scenario the benefit would derive from the ability of the arg/lys regions in a heterodimeric protein pair to recognize two different nucleotide sequences within a single DNA motif. That ability would expand an organism's repertoire of usable regulatory motifs, because the presence of the heterodimers would enable cells to exploit asymmetric dyad motifs, which in contrast to dyad-symmetric motifs have nonidentical halves [*see left side of illustration on opposite page*].

The increase in the repertoire would arise, however, only if homodimers were actually restricted to recognizing dyad-symmetric motifs. As it turns out, my co-workers and I have found preliminary evidence that homodimers and heterodimers bind asymmetric motifs equally well. Our methods for studying DNA binding may not be subtle enough to mimic conditions within the living cell. Nevertheless, if our initial results are confirmed, they would indicate that the special value of heterodimers lies elsewhere.

Alternatively, heterodimers might serve to increase the number of different jigsaw puzzles that can be formed by a limited set of regulatory proteins. For this notion to make sense, it is necessary to discuss yet another functional domain that has been found in gene regulatory proteins.

As Ptashne and his colleague Ann Hochschild first recognized, gene activation by proteins bound to DNA may be achieved by parts of regulatory proteins that are quite distinct from the parts responsible for binding DNA. In studies of the cI protein encoded by the bacteriophage lambda (a virus that infects bacteria), the Harvard workers showed that a region critical for gene activation was totally dispensable for DNA binding. They predict that this cI activation domain interacts in some way with the RNA polymerase that carries out transcription.

Leucine zipper proteins, including C/EBP, Fos, Jun and GCN4, also have functional domains that are distinct from their zipper and DNA-binding segments. By analogy, they too can be called activation domains. The biochemical role played by these regions is not fully understood, but they may well serve as docking sites for other proteins. Even within families of cross-dimerizing proteins, these domains will differ. Hence, two varieties of C/EBP that can zipper together will each have a distinct activation domain, as will Fos and Jun, which, it will be recalled, combine with each other.

Heterodimers of leucine zipper proteins, then, could expand the number of distinguishable jigsaw puzzles available to an organism because the activation domains of the dimer subunits would each bind to a different protein. In contrast, if only homodimers existed, the proteins contacted by a DNA-bound dimer would always have to be identical [*see right side of illustration on opposite page*].

The potential importance of heterodimers is underscored by the finding that cross-mixing probably occurs in many classes of gene regulatory proteins, not only in those with leucine zippers. Indeed, Harold Weintraub of the Fred Hutchinson Cancer Research Center in Seattle and David Baltimore, now at the Rockefeller University, have identified cross-mixing (by a means other than the leucine zipper) in an entirely different class of gene regulators, known as helix-loop-helix (HLH) proteins. Weintraub and his colleague Andrew B. Lasser have shown that one such protein, called MyoD, is central to the specialization of muscle cells.

Although I have concentrated on leucine zipper proteins in this article, I must emphasize that many different classes of gene regulatory proteins participate in gene activation and the specialization of cells. For example, James E. Darnell, Jr., and his colleagues at Rockefeller have found that a number of regulatory proteins from different classes give rise to the specialized activities of differentiated hepatocytes.

Those of us interested in better understanding cell specialization must now determine how the many different regulatory proteins in cells fit together on their DNA templates and communicate with one another. The discovery of the leucine zipper has solved part of the problem. No one knows when the full answer will emerge, but the problem finally begins to seem conquerable.

FURTHER READING

THE LEUCINE ZIPPER: A HYPOTHETICAL STRUCTURE COMMON TO A NEW CLASS OF DNA BINDING PROTEINS. W. H. Landschulz, P. F. Johnson and S. L. McKnight in *Science*, Vol. 240, pages 1759–1764; June 24, 1988.

EUKARYOTIC TRANSCRIPTIONAL REGULATORY PROTEINS. Peter F. Johnson and Steven L. McKnight in *Annual Review of Biochemistry*, Vol. 58, pages 799–839; 1989.

EVIDENCE THAT THE LEUCINE ZIPPER IS A COILED COIL. E. K. O'Shea, R. Rutkowski and P. S. Kim in *Science*, Vol. 243, pages 538–542; January 27, 1989.

THE DNA BINDING DOMAIN OF THE RAT LIVER NUCLEAR PROTEIN C/EBP IS BIPARTITE. W. H. Landschulz, P. F. Johnson and S. L. McKnight in *Science*, Vol. 243, pages 1681–1688; March 31, 1989.

ACTION OF LEUCINE ZIPPERS. Ted Abel and Tom Maniatis in *Nature*, Vol. 341, No. 6237, pages 24–25; September 7, 1989.

SCISSORS-GRIP MODEL FOR DNA RECOGNITION BY A FAMILY OF LEUCINE ZIPPER PROTEINS. C. R. Vinson, P. B. Sigler and S. L. McKnight in *Science*, Vol. 246, pages 911–916; November 17, 1989.

Collisionless Shock Waves

To many theorists' surprise, shock waves form even in the rarefied material between the planets. These shocks sculpt the space environment and lie at the heart of a variety of astrophysical phenomena

by Roald Z. Sagdeev and Charles F. Kennel

Shock waves resonate through the solar system, much like the reverberating boom from a supersonic jet. In the latter case, the disturbance is caused by an aerodynamic shock, an abrupt change in gas properties that propagates faster than the speed of sound. It had long been recognized that in a neutral gas, such as the earth's atmosphere, particles must collide if shocks are to form. Beginning in the 1950s, we and our colleagues theorized that, contrary to the expectations of many scientists, similar shock waves could form even in the near vacuum of outer space, where particle collisions are extremely rare. If so, shocks could play a significant role in shaping space environments.

"Collisionless" shocks cannot occur naturally on the earth, because nearly all matter here consists of electrically neutral atoms and molecules. In space, however, high temperatures and ultraviolet radiation from hot stars decompose atoms into their constituent nuclei and electrons, producing a soup of electrically charged particles known as a plasma. Plasma physicists proposed that the collective electrical and magnetic properties of plasmas could produce interactions that take the place of collisions and permit shocks to form.

In 1964 the theoretical work found its first experimental confirmation. Norman F. Ness and his colleagues at the Goddard Space Flight Center, using data collected from the *IMP-1* spacecraft, detected clear signs that a collisionless shock exists where the solar wind encounters the earth's magnetic field. (Solar wind is the continuous flow of charged particles outward from the sun.) More recent research has demonstrated that collisionless shocks appear in a dazzling array of astronomical settings. For example, shocks have been found in the solar wind upstream (sunward) of all the planets and comets that have been visited by spacecraft. Violent flares on the sun generate shocks that propagate to the far reaches of the solar system; tremendous galactic outbursts create disruptions in the intergalactic medium that are trillions of times larger. In addition, many astrophysicists think that shocks from supernova explosions in our galaxy accelerate cosmic rays, a class of extraordinarily energetic elementary particles and atomic nuclei that rain down on the earth from all directions.

The study of plasmas began in the 19th century, when Michael Faraday investigated electrical discharges through gases. Modern plasma research dates from 1957 and 1958. During those years, Soviet *Sputnik* and American *Explorer* spacecrafts discovered that space near the earth is filled with plasma. At the same time, till then secret research on controlled thermonuclear fusion conducted by the U.S., Soviet Union and Europe was revealed at the Atoms for Peace Conference in Geneva, greatly increasing the freely available information on plasmas.

Fusion research focuses on producing extremely hot plasmas and confining them in magnetic "bottles," to create the conditions necessary for energy-producing nuclear reactions to occur. In 1957, while searching for a method to heat fusion plasmas, one of us (Sagdeev) realized that an instantaneous magnetic compression could propagate through a collisionless plasma, much as a shock moves through an ordinary fluid.

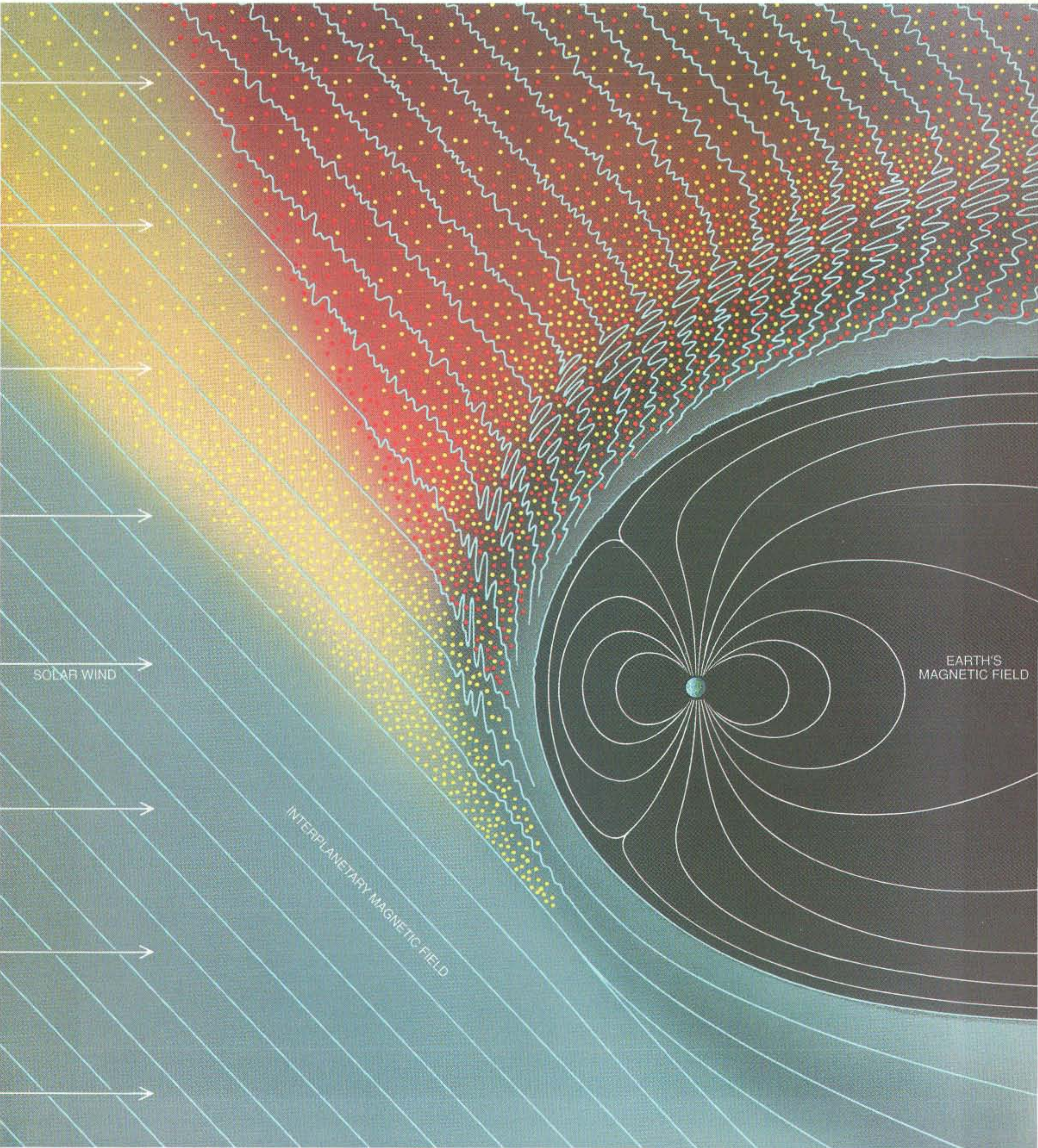
Magnetic fields that thread through plasmas make them behave somewhat like such a fluid. A magnetic field exerts a force (the Lorentz force) on a moving electrically charged particle. The field can be thought of as a series of magnetic lines through the plasma, like the field lines around a bar magnet that can be made visible with iron filings. The Lorentz force always acts perpendicular both to the direction of the magnetic field line and to the direction in which a particle is moving. If the particle moves perpendicular to the field, the force acts like a rubber band, pulling the particle back and constraining it to move in small circles about the magnetic field line. The particle can, however, move freely in the direction of the magnetic field line. The combination of the free motion along and constrained, circular rotation across the magnetic field shapes the particle's trajectory into a helix that winds around a magnetic field line.

The Lorentz force makes it difficult to disperse the plasma in the direction

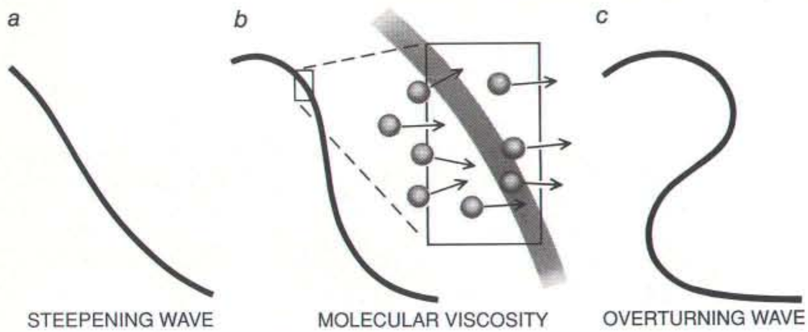
ROALD Z. SAGDEEV and CHARLES F. KENNEL have been friends and colleagues since they met at the International Centre for Theoretical Physics in Trieste in 1965. Sagdeev heads the theory division of the Soviet Institute of Space Research and is professor of physics at Moscow Physico-Technical Institute. Last year he joined the physics department of the University of Maryland at College Park. In addition to his astronomical and physical research, Sagdeev has been active in the areas of arms control, science policy and global environment protection. Kennel is professor of physics at the University of California, Los Angeles, as well as a consultant to TRW Systems Group, where he participates in space plasma experiments. He is also a distinguished visiting scientist at the Geophysical Institute of the University of Alaska, Fairbanks, and a collector of native Alaskan art.

COLLISIONLESS SHOCK forms as the solar wind, a stream of charged particles flowing from the sun, hits the earth's magnetic field. The associated phenomena depend on the direction of the upstream interplanetary magnetic field. Energetic electrons (*yellow*) and ions (*red*) escape upstream in the quasiparallel zone (*top regions*). The escaping particles create microscopic waves in the solar wind and torsional magnetic disturbances, known as Alfven waves (*wavy lines*). The shock is thin where it is quasiperpendicular (*bottom regions*).

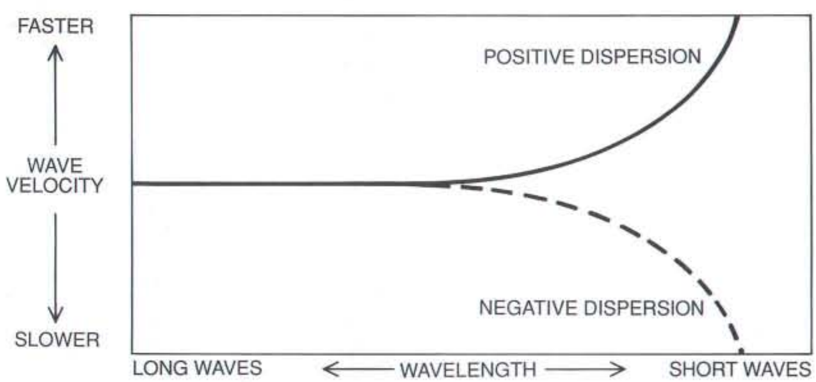
perpendicular to the magnetic field. The maximum distance over which particles can move away from the field, called the Larmor radius, is inversely proportional to the field strength. In the weak interplanetary magnetic field, the Larmor radius amounts to several kilometers for electrons and several hundred kilometers for more massive ions. These distances may seem large, but they are tiny compared with the size of the region where the solar wind encounters the earth's magnetic field. The shock that forms there, called a bow shock, has the same parabolic shape as the waves that pile up ahead of a speedboat. It stretches more than 100,000 kilometers across. When the scale is larger than the Larmor radius for ions, the collective motion of plasma particles across the magnetic field



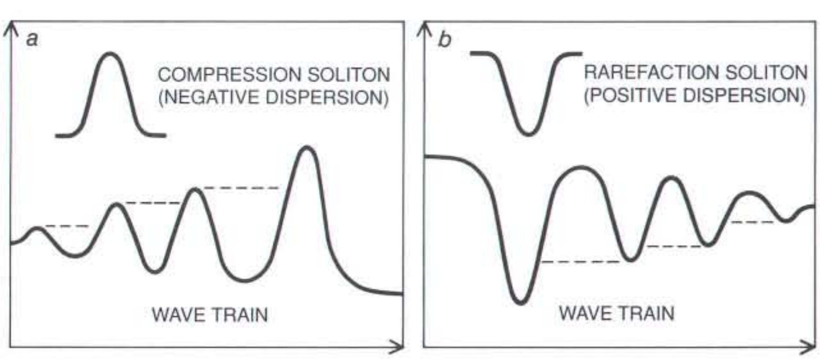
How Collisionless Shocks Form



Waves steepen because the higher-amplitude part of a wave moves faster. As time passes, a wave (*a*) grows steeper and more abrupt. In a sound wave, collisions between individual air molecules (molecular viscosity) transfer momentum forward (*b*) and prevent the wave from overtaking the slow-moving material that lies upstream. In strong collisionless shocks, this process can lead to overturning of the wave (*c*).



Wave dispersion dictates that a wave's velocity depends on its wavelength. This effect becomes significant for collisionless shocks in plasmas (gases of electrons and bare nuclei). The figure shows that wave speed can increase (positive dispersion) or decrease (negative dispersion) with increasing wavelength. Dispersion tends to smudge a sharply defined wave into a long series of its component harmonic waves.



Solitary waves, or solitons, form through the competing actions of steepening and dispersion. Depending on the sign of the dispersion, solitons can be regions of compression (*a*) or pulses of rarefaction (*b*). An ordered sequence of solitons, known as a wave train, defines the thickness of a shock front. The length of the wave train depends on the damping mechanism that causes solitons to lose energy and ultimately to dissipate into heat.

actually drags the field lines along with it. The magnetic field thus becomes "frozen" into the plasma.

In short, a magnetic field endows collisionless plasmas with elastic properties analogous to those of a dense gas, and so a plasma wave crossing a magnetic field behaves somewhat like an ordinary sound wave. The theoretical analysis of collisionless shocks therefore started by following the ideas developed from earlier research on aerodynamic shocks.

Suppose, for example, a sudden compression creates a sound wave in air. As the wave travels, its shape—that is, its profile of pressure and density—changes. Because the most compressed regions of the wave move the fastest, the wave grows stronger and its leading edge becomes sharper. The great German mathematician Bernhard Riemann showed how this phenomenon, called wave steepening, creates shock waves.

Ultimately the faster-moving denser air behind catches up with the slower air ahead. At this point, the sound wave behaves somewhat like an ocean wave heading toward shore. A water wave steepens, overturns and then crashes into foam. A sound wave reaches an analogous but different climax. As the wave grows so steep that it is about to overturn, individual gas molecules become important in transporting momentum between neighboring points in the gas: molecules from the faster, denser region of the wave rush ahead of the steepening wave front, colliding with molecules in the slower region ahead of the wave and exchanging momentum with them. In this way, the slower molecules are brought up to the speed of the moving wave.

This exchange of momentum is caused by molecular viscosity. In this process, momentum is passed from the overtaking wave crest and imparted to the undisturbed region ahead of it, much as in a relay race a baton is passed from one runner to the next. Molecular viscosity becomes highly efficient when the thickness of the wave front shrinks to the average distance that a particle can travel before it collides with another, a distance known as the collision mean free path. (The mean free path of a molecule in air is about one ten-thousandth of a centimeter long.) At this thickness, steepening and viscosity balance each other, and a steady shock wave forms. The resulting shock represents an almost steplike change in gas velocity, density and pressure.

Before physicists knew of a mechanism that could replace molecular viscosity in plasmas, it made little sense

for them to talk of collisionless shocks. Consequently, the topic lay fairly dormant for many years. Then, in the late 1950s, one of us (Sagdeev) and, independently, Arthur R. Kantrowitz and Harry E. Petschek, then at the Avco-Everett Research Laboratory near Boston, suggested that a similar sort of momentum relay race could take place in a tenuous plasma. They theorized that in a plasma, waves rather than individual particles pass along the baton.

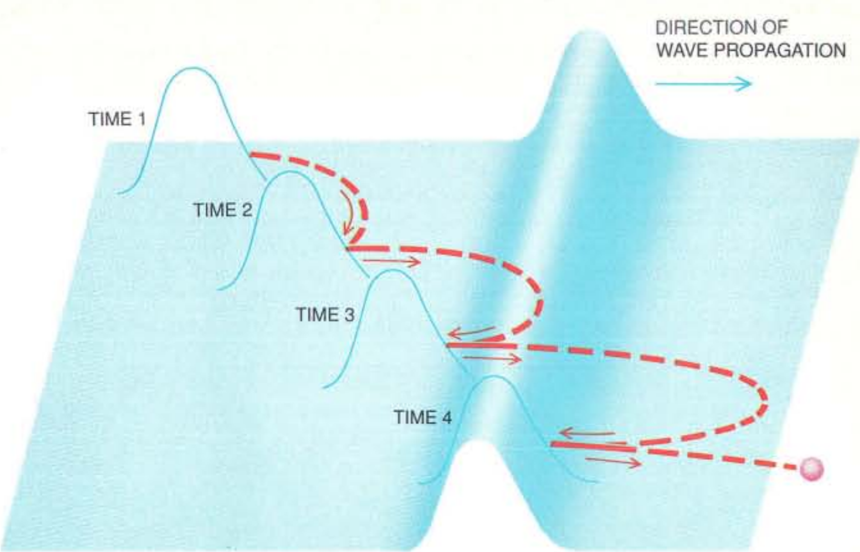
The plasma relay race depends on the fact that the speed of a plasma wave changes with wavelength, an effect called dispersion. Indeed, whereas in ordinary gases the speed of a sound wave is practically independent of wavelength, in collisionless plasma a wave is very dispersive. That is, its speed may either increase or decrease as its wavelength shortens, depending on the angle between the direction of propagation of the wave and the orientation of the magnetic field.

According to Fourier's theorem, a fundamental theorem of mathematics, any wave profile consists of many superimposed waves, or harmonics, of different wavelengths. (By analogy, white light is composed of many distinct colors, each of a different wavelength.) If the wave profile steepens, it excites harmonics of ever shorter wavelength.

For wave propagation that is not exactly perpendicular to the magnetic field, dispersion causes shorter-wavelength harmonics to travel faster than the longer-wavelength ones (negative disperson). The effects of dispersion become significant when a steepening shock front becomes about as thin as the Larmor radius for ions. At this point, the shorter-wavelength harmonics race ahead of the front into the undisturbed plasma upstream. These harmonics carry along the momentum, like the fast molecules in a sound wave.

The competing actions of steepening and dispersion yield a series of wave pulses that propagate in the direction of the shock. As a result, the front acquires the shape of a "wave train." The weakest (smallest-amplitude) waves announce the arrival of the train, and successively stronger oscillations build up until the full shock transition arrives. The length of the train (in other words, the thickness of the shock front) depends on how rapidly the energy of the waves dissipates.

For waves propagating exactly perpendicular to the magnetic field, dispersion causes the harmonic wave speed to decrease at shorter wavelengths. Short-wavelength harmonics now trail behind the shock front, and so they cannot affect steepening of the overall wave. In this case, the shock passes the momentum baton to a series of compressional pulses called solitons.



WAVE DAMPING occurs as energy is transferred from solitons to individual charged particles (electrons or ions) in the plasma. Particles traveling at roughly the velocity of the plasma wave interact strongly with the solitons. The solitons' electric fields force these resonant particles to stop and reverse course, which drains energy from the solitons and imparts it to the particles. The restoring magnetic force (the Lorentz force) curves each particle's path, returning it again and again to the same soliton. Each encounter boosts the particle's energy at the expense of the soliton.

Solitons in perpendicular shocks are approximately the thickness of an electron's Larmor radius, and they are created when the wave profile steepens to that scale. The steepening front radiates an ordered sequence of solitons, led by the largest (highest-amplitude) one and trailed by successively smaller ones that ultimately blend into the smooth state behind the shock. The length of the soliton train depends on how fast the soliton energy is dissipated into heat.

Waves on the surface of shallow water behave very much like dispersive waves in collisionless plasma. The theory of shallow water waves was developed in the late 19th century, culminating in the classic work of Diederik J. Korteweg and G. DeVries that first described the solitons that occasionally propagate down Dutch canals. The seemingly recondite analogy between shallow water solitons and plasma solitons expresses a general physical truth: solitons can form whenever wave steepening and dispersion compete.

One implication of this fact is that solitons form even in shocks that do not propagate exactly perpendicular to the magnetic field. The wave pulses mentioned earlier can also be thought of as solitons, the difference being that these solitons are rarefactive (low density) rather than compressive. In this case, short-wavelength harmonics travel relatively slowly (positive disperson), and the greater the amplitude of the rarefactive soliton, the more slowly it propagates. As a result, the wave train terminates with the strongest soliton. Surface tension in water creates small waves that have positive dispersion and rarefactive solitons. The physics of water waves therefore provides an analogy to both types of dispersion found in collisionless plasma.

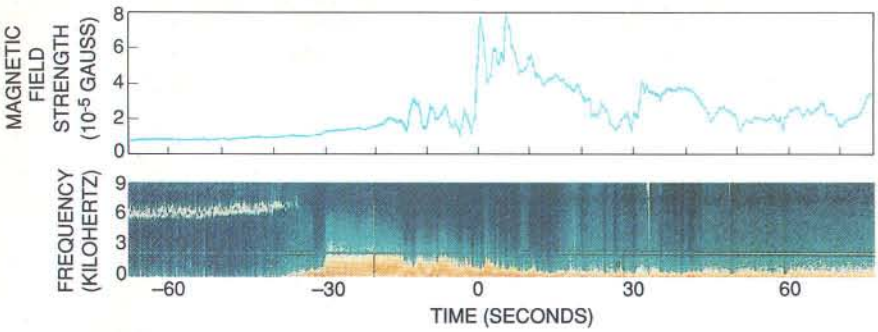
The elegant theory of solitons is an impressive achievement of modern mathematical physics. In 1967 Martin Kruskal and his colleagues at Princeton University proved that any wave profile in a dispersive medium that can support steepening evolves into a sequence of solitons. By relating soliton theory to the problem of elementary particle collisions, which has been studied in depth in quantum physics since the 1920s, they showed that solitons preserve their identities when they collide, just as particles do.

The understanding of dispersive shocks remains incomplete without a knowledge of how to dissipate the energy of waves or solitons into heat. If not for the effect of dissipation, the train of wave structures making up the shock front would be infinitely long. In effect, the fundamental question of how collisionless shock waves transport energy and momentum has reappeared, but in a new guise.

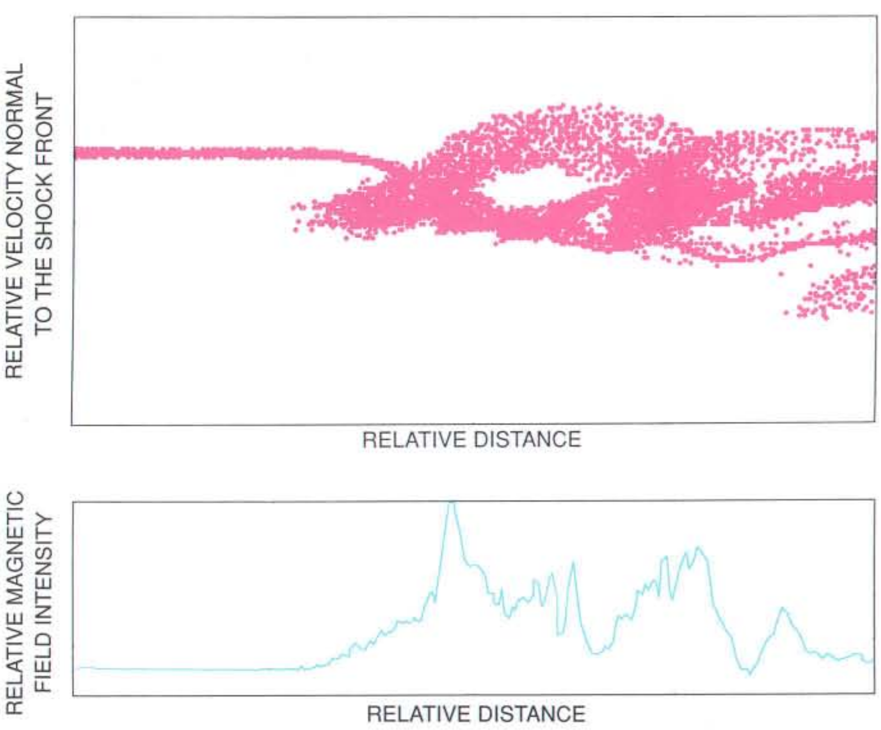
In 1945 the great Soviet physicist Lev

D. Landau discovered a dissipation mechanism that requires no collisions between particles. Among the randomly moving particles in a plasma, a few happen to travel at a velocity that matches the velocity of the plasma wave. These particles are said to be in resonance with the wave. An intense exchange of energy can take place between a wave and the particles resonant with it.

In the early 1970s one of us (Sagdeev) and Vitaly Shapiro, also at the Institute of Space Research in Moscow, showed that Landau's mechanism damps solitons by accelerating resonant ions. Consider, for example, a train of compressive solitons propagating perpendicular to the magnetic field. Each soliton generates an electric field parallel to its direction of motion. Ions traveling close to the resonant velocity move slowly compared with the solitons, and the soliton electric field is able to stop and reverse the motion of these ions. The soliton loses part of its energy to the ions resonant with it during the interaction.



MAGNETIC OSCILLATIONS (*top*) were recorded by *Voyager 1* as it crossed the collisionless shock in the solar wind ahead of Jupiter. Upstream is to the left. Magnetic disturbances preceding the shock front were caused by ions reflected back into the solar wind. Measurements of microscopic plasma waves (*bottom*) showed that the reflected ions created low-frequency ion acoustic waves. Electrons heated by ion waves escaped upstream, generating higher-frequency electron oscillations.



COMPUTER SIMULATION clarifies the behavior of ions at the Jovian shock (*top*). Each dot represents a particle having a certain position and velocity at one instant in a steepening, breaking and overturning shock. Streams of "solar wind" ions (*to the left*) reflect off the shock front while those farther downstream (*right*) move chaotically. The large peak and subsequent oscillations of the magnetic field seen in the simulation (*bottom*) resemble those of the real shock. Dennis Papadopoulos and Peter Cargill of the University of Maryland at College Park ran the simulation.

The process does not end here, because the magnetic Lorentz force curves the path of the reflected ion so that it returns again and again to the same soliton. Each encounter adds to the energy of the particle. The Lorentz force, which grows stronger as the particle velocity increases, eventually throws the ion over the top of the first soliton. The acceleration continues as the ion encounters the remaining solitons in the wave train. The resonant ions gain energy much as surfers gain speed by riding ocean waves. This analogy inspired John M. Dawson of the University of California at Los Angeles to design a new kind of charged particle accelerator, which he dubbed the surfatron.

The heating of ions by solitons can form a shock if the number of ions in resonance is great enough. Such is the case if the ions are hot. If not, the solitons find another way to dissipate energy: they themselves generate microscopic plasma waves that heat the plasma. Plasma electrons flow over ions, thereby creating the electric current responsible for the characteristic soliton magnetic field profile. If the ions are cold, the electrons can easily move at supersonic velocities relative to the ions, in which case the electrons amplify extremely small scale electric field oscillations called ion acoustic waves. These waves, which do not affect the magnetic field, grow in an avalanchelike fashion. The plasma particles collide not with one another but with these ion acoustic waves. After the waves develop, the plasma enters a microturbulent state.

In 1968 Robert W. Fredericks and his colleagues at TRW in Los Angeles were the first to detect ion acoustic waves in shocks. They made this discovery using instruments on the *OGO-5* spacecraft that were designed specifically to study plasma waves in space. Since then, plasma-wave detectors have been included on most space missions concerned with solar system plasmas, notably the *International Sun-Earth Explorers* (*ISEE 1, 2* and *3*) in earth orbit and the *Voyager 1* and *2* missions to the outer planets. The late Fred Scarf of TRW and his collaborators often played back the microturbulent-wave electric fields recorded by the *ISEE* and *Voyager* spacecraft through an ordinary loudspeaker. To most listeners, shocks would sound cacophonous; to our ears, however, they were a symphony of space.

Although easy to record, microturbulence has proved difficult to understand completely. Theorists turned to numerical computations to help elucidate the behavior of a strongly microturbulent plasma. By solving millions of equations of motion for the particles, computer simulation shows how ion acoustic waves grow and heat the plasma. Today's supercomputers are just beginning to give scientists comprehensive understanding of many different kinds of microturbulence.

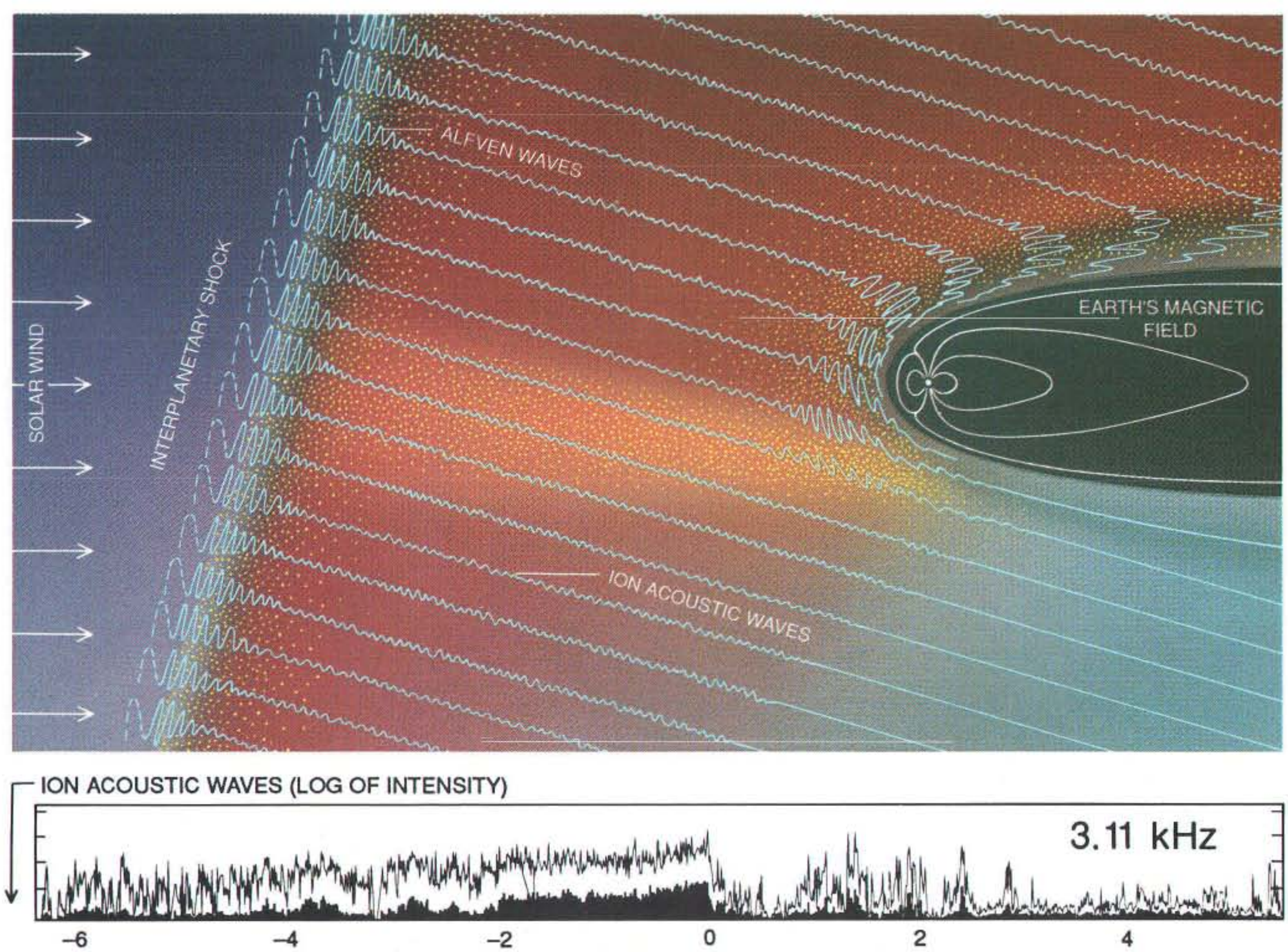
Even without knowing the detailed nature of microturbulent plasma, physicists can deduce its general behavior. Electrons in the plasma transfer their momentum to ion acoustic waves, which in turn transfer it to ions. This process retards the motion of the electrons in the plasma and so creates resistance to the electric current. In some shocks, ion acoustic-wave resistance grows sufficiently intense to suppress the generation of solitons. When this happens, no wave train forms, and the shock is called resistive.

Although both simple dispersive and resistive shocks have been found in space, most shocks observed there have entirely different characteristics from those discussed so far. Most shocks are sufficiently powerful that neither dispersion nor resistance can prevent steepening from causing the waves to overturn. Overturning then leads to a host of new shock phenomena.

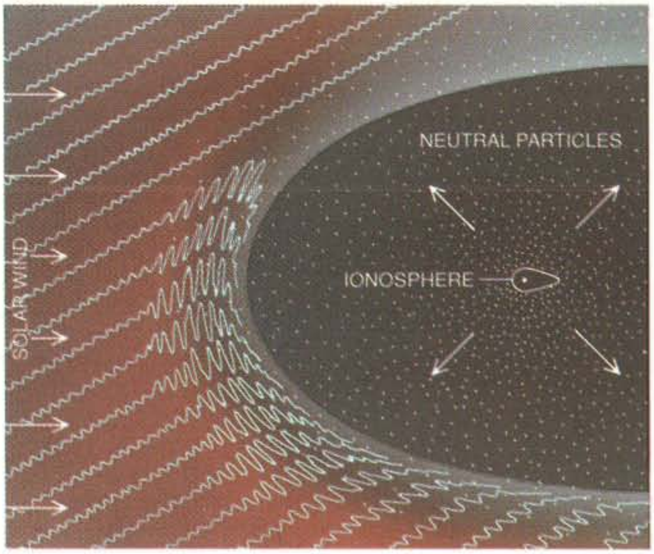
A consideration of shallow water waves, once more, helps to illustrate the process of overturning. When a shallow ocean wave grows sufficiently high, the tip of its wave crest swings forward through an arc and ultimately collapses under gravity. The water stream from behind the crest collides with that ahead, giving rise to the foam on whitecaps. Thus, a large wave crashing toward shore repeatedly overturns, or "breaks."

A plasma wave also develops overlapping velocity streams as it overturns. The fastest stream, which comes from the wave crest, invades the plasma ahead of the shock front. The Lorentz force turns the ions in this stream back into the shock. These reflected ions ultimately mix with those behind the front. If the shock is weak, its structure will remain steady. If the shock is strong, ion reflection will temporarily overwhelm steepening; however, the shock soon steepens again, and the cycle repeats. Recent numerical simulations by Kevin B. Quest and his colleagues at Los Alamos National Laboratory confirm the idea that very strong shock waves consist of a repeated cycle of steepening, overturning and ion reflection.

The interactions between reflected and flowing ions can also lead to microturbulence. The *Voyager* spacecraft detected ion acoustic waves, this time generated by ions reflected by Jupiter's



INTERPLANETARY SHOCK begins with a solar eruption and travels billions of kilometers through the solar wind. Instruments on the *ISEE-3* spacecraft measured the passage of a large quasiparallel interplanetary shock just before it swept over the earth. The features of the earth's bow shock are superimposed on similar ones related to the far larger interplanetary shock (*top*). Microscopic plasma waves (*bottom*), along with magnetic Alfven waves and accelerated particles, reached *ISEE-3* several hours ahead of the interplanetary shock. This turbulent region was about one million kilometers thick.



COMETARY SHOCKS result when neutral molecules and atoms from a comet's nucleus escape into the solar wind before they are ionized by ultraviolet radiation. The addition of heavy cometary ions slows the flow of the solar wind (because of momentum conservation) and thereby produces a shock. Because the heavy ions generate extremely intense Alfven waves, cometary shocks are thicker and more turbulent and involve more particle acceleration than do shocks around planets (*left*). The shock is far larger than the comet's nucleus and even than the huge visible cometary tail (*right*).

bow shock [*see top illustration on page 44*]. Near the earth, reflected ions generate waves in the solar wind at the geometric mean of the frequencies of rotation of the ions and electrons about the earth's magnetic field; this mean is called the lower hybrid-resonance frequency. In 1985 the Soviet-Czech *Intershock* spacecraft made the first definitive measurements of lower hybrid turbulence in the earth's bow shock. Around both planets, the ion acoustic waves take energy from ions and give it to electrons. Some heated electrons escape forward into the solar-wind flow, others back into the shock zone.

So far we have concentrated on those shocks propagating more or less at right angles to the magnetic field, those physicists call quasiperpendicular. Plasma turbulence is even more important when the shock propagates almost parallel to the magnetic field. The field no longer holds back the fast particles that rush ahead of a quasiparallel shock. These particles are a major source of turbulent instability.

The ability of the magnetic field to channel particle motion along field lines creates a situation analogous to a fire hose left spraying water on the ground. Bends in the hose become increasingly curved by the centrifugal force of the flowing water; eventually the hose wriggles uncontrollably on the ground.

The magnetic field channeling the overlapping plasma streams ahead of a quasiparallel shock experiences a similar instability, often called the fire-hose instability. The centrifugal force that bends the magnetic field lines is proportional to the density of energy in plasma motion along the magnetic field. Instability occurs when this energy density exceeds that of the magnetic field. Many physicists conceived of the fire-hose instability independently, but the version invented in 1961 by Eugene N. Parker of the University of Chicago was tailored specifically to quasiparallel shocks.

The plasma fire-hose instability leads to a random flexing of the magnetic field lines. This kind of magnetic turbulence can be regarded as a chaotic ensemble of "torsional" waves, that is, ones that twist the magnetic field lines. They are known as Alfven waves, after Hannes Alfven of the Royal Institute of Technology in Stockholm, who first described them.

Alfven waves, like ion acoustic waves, can exchange energy and momentum with ions in resonance with them. As far as the ions are concerned, the interaction with Alfven waves mimics the effect of collisions. Thus, Alfven waves limit how far ions escaping the shock can penetrate upstream and determine the thickness of the quasiparallel shock.

Theory predicts that collisions between ions and Alfven waves should be nearly elastic, that is, they should involve only slight changes in energy despite a large change in momentum (for example, when a rubber ball bounces off a hard wall, its momentum reverses, but its energy remains essentially the same). As a result, the Alfven turbulence inside the shock front should disintegrate relatively slowly. This notion led us to conclude in 1967 that quasiparallel shocks could be much thicker than quasiperpendicular ones.

The very first measurements of the earth's bow shock by the *IMP-1* spacecraft in 1964 hinted at the substantial differences between parallel and perpendicular shocks. The data returned by *IMP-1* were somewhat puzzling at first because sometimes the shock appeared thin and other times it appeared thick. Three years later we suggested that shock structure could depend on the orientation of the interplanetary magnetic field. In 1971 Eugene W. Greenstadt of TRW and his colleagues assembled the first evidence that the thickness of the earth's bow shock does indeed vary with the direction of the solar-wind magnetic field. Since this field constantly changes direction, the regions where the bow shock is locally quasiperpendicular and where it is quasiparallel are always moving, even if the shock itself remains fairly stationary. Wherever the shock is quasiperpendicular, it is thin; where it is quasiparallel, it is thick [*see illustration on page 41*].

In the early 1970s spacecraft began to detect small fluxes of energetic particles, ion acoustic waves and Alfven waves far upstream of where the earth's bow shock was understood to be. The ISEE program, which started in 1977, established that all the upstream activity is actually part of the extend-

ed quasiparallel shock. The shock is so thick that it dwarfs the earth, and therefore earth-orbiting satellites cannot really measure its size.

Another, larger class of shocks does lend itself to investigation by spacecraft, however. Flares in the solar corona occasionally launch gigantic shock waves that propagate through the interplanetary medium to the far reaches of the solar system. These can be observed as they sweep by instrumented spacecraft. One of us (Kennel), along with colleagues in the ISEE project, found that the region of Alfven and ion acoustic turbulence upstream of quasiparallel interplanetary shocks can be more than a million kilometers thick.

Alfven waves play a particularly prominent role in the shocks that form ahead of comets as they pass through the solar wind in the inner solar system. Cometary nuclei are far too small to cause any detectable physical disturbance in the flow of the solar wind (the nucleus of Halley's comet, for instance, measures only about 15 kilometers across), and the nuclei possess a negligible magnetic field. Because of these properties, comets cannot generate shocks in the way that planets do. Nevertheless, scientists have found that when comets approach the sun, they create large collisionless shocks.

Sunlight evaporates atoms and molecules from the surface of a comet's nucleus. Most of the liberated gas is ionized by solar ultraviolet light and forms a plasma cloud similar to the earth's ionosphere. The solar wind never penetrates the cometary ionosphere, and it is not the ionosphere that forms the shock wave. The key players in producing cometary shocks are the few neutral atoms and molecules that manage to escape the comet's ionosphere. These, too, are ultimately ionized, but farther out, where they have entered the solar wind.

The newly ionized particles respond to the electric and magnetic fields of the solar wind by joining the flow. They increase the mass density of the solar wind, which, according to the law of conservation of momentum, decreases the wind speed. Because cometary ions are much heavier than the protons of the solar wind, a number of cometary ions can slow the wind appreciably.

More than 20 years ago Ludwig Biermann of the Max Planck Institute for Astrophysics in Munich suggested that such a decelerating solar-wind flow should produce a shock similar to a planetary bow shock. During its 1986 encounter with Comet Halley, the Soviet spacecraft *Vega-1* heard the plasma wave cacophony that signaled the existence of a shock wave about one million kilometers from the nucleus, the distance predicted by Biermann's theory.

The Soviet *Vega*, Japanese *Suisei* and the European *Giotto* spacecraft encountered both quasiperpendicular and quasiparallel shocks at Comet Halley. The quasiparallel shocks were similar to those at the planets. Heavy ions upstream of the quasiperpendicular cometary shocks generated intense Alfven-wave turbulence, however, something that does not happen around the planets.

Shocks that generate Alfven waves can also accelerate a small group of particles to high energies. The "collisions" of particles with Alfven waves return escaping particles back to the shock front. Each time they recross the shock, the particles increase their energy. This acceleration mechanism is based on one proposed by Enrico Fermi in 1954. In 1986 one of us (Kennel) and his ISEE collaborators found that a theory of Fermi acceleration developed for interplanetary shocks by Martin A. Lee of the University of New Hampshire successfully passed the test of observations. Yet the Fermi process develops so slowly that the protons accelerated by quasiparallel interplanetary shocks only reach energies of a few hundred thousand electron volts in the one day it takes the shock to travel from the sun to the earth. In comparison, cosmic rays—energetic subatomic particles and atomic nuclei from deep space—have energies up to 100 trillion electron volts.

Exploding stars—supernovas—create very strong shocks that speed into the interstellar plasma at tens of thousands of kilometers per second. We cannot put a space probe ahead of a supernova shock, so we cannot say for sure whether the shock generates Alfven waves and accelerates interstellar ions. We can, however, apply to supernova shocks the theory of particle acceleration that is being tested today using solar system shocks.

Since supernova shocks last about a million years before dying out, particles have time to reach extremely high energies via the Fermi process. Working independently, Germogen F. Krymskii of the Institute of Space Physics Research and Aeronomy in Yakutsk, U.S.S.R., Roger D. Blandford of the California Institute of Technology and Ian W. Axford of the Max Planck Institute for Aeronomy in Katlenburg-Lindau, together with their colleagues, showed in 1977 that the distribution in energy of the particles accelerated by collisionless shocks is virtually identical to that of cosmic rays.

The origin of cosmic rays has long been a puzzle. Many astrophysicists now believe that they are created when supernova shocks accelerate particles, although it is still not understood how the particles reach the highest energies observed.

Collisionless shocks probably exist even around remote galaxies. Dynamic processes in the centers of some active galaxies (possibly involving a massive black hole) create supersonic jets hundreds of thousands of light-years long. Shocks are likely to occur when the jets interact with the plasma surrounding the galaxy. Radio emissions from the jets indicate that electrons are accelerated to extremely high energies. Albert A. Galeev, director of the Soviet Institute of Space Research, suggests that a theory he and his colleagues developed to explain how lower hybrid waves accelerate electrons in the earth's bow shock may also clarify how electrons are accelerated in galactic jets.

Contemporary collisionless shock research encompasses phenomena that vary tremendously in scale and origin. The concepts that we and others developed 20 years ago have turned out to be a reasonable basis for understanding collisionless shocks. Spacecraft have found individual examples of most of the shock types predicted by theory. Still to come are refined measurements and numerical calculations that simulate in detail the impressive variety of shocks found in nature. In most cases, the fairly simple mechanisms we have described here are intertwined in fascinating ways. Yet even now collisionless shock theory has enabled physicists to speculate with some confidence on the physical processes underlying some of the grandest and most violent phenomena in the universe.

FURTHER READING

SHOCK WAVES IN COLLISIONLESS PLASMAS. D. A. Tidman and N. A. Krall. Wiley-Interscience, 1971.

UPSTREAM WAVES AND PARTICLES. *Journal of Geophysical Research*, Vol. 86, No. A6, pages 4319–4529; June 1, 1981.

HANDBOOK OF PLASMA PHYSICS. Edited by M. N. Rosenbluth and R. Z. Sagdeev. North-Holland Publishing Company, 1983.

COLLISIONLESS SHOCKS IN THE HELIOSPHERE: REVIEWS OF CURRENT RESEARCH. Edited by Bruce T. Tsurutani and Robert G. Stone. American Geophysical Union, 1985.

NONLINEAR PHYSICS: FROM THE PENDULUM TO TURBULENCE AND CHAOS. R. Z. Sagdeev, D. A. Usikov and G. M. Zaslavsky. Translated from the Russian by Igor R. Sagdeev. Harwood Academic Publishers, 1988.

Hepatitis B Virus

This small, extraordinary virus causes liver diseases and a common form of cancer. New vaccines produced by genetic engineering hold the promise of eventually eradicating both

by Pierre Tiollais and Marie-Annick Buendia

The liver disease called hepatitis B is a widespread and serious health problem in its own right, but it conceals a greater menace: the virus that causes hepatitis B is second in importance only to tobacco as a known human carcinogen. Hundreds of millions of people, most of them in regions with poor medical care, are chronically infected with the virus and face an elevated risk of acquiring liver cancer. Moreover, although many chronic carriers appear healthy, they can still transmit the hepatitis B virus to those with whom they have close contact, thereby starting the cycle of disease anew.

Fortunately, the prospects for interrupting that cycle have brightened considerably during the past decade. Recombinant DNA technology, or genetic engineering, has unlocked many of the secrets of the hepatitis B virus (HBV). Today researchers understand the peculiar life history of HBV and are probably closing in on answers to how HBV causes cancer. Moreover, vaccines produced by genetic engineering can prevent HBV from spreading.

The first step in the identification of the hepatitis B virus was made in 1963, when Baruch S. Blumberg, then at the Institute for Cancer Research in Philadelphia, was studying certain proteins in blood serum. In a sample from a patient with hemophilia, he observed an antibody that reacted with an antigen present in the blood of an Australian aborigine infected with hepatitis. In 1968 Blumberg identified it as the HBV surface antigen (HBsAg).

For the next decade, the absence of cell-culture systems for propagating HBV hampered investigators. Then in 1978 our laboratory and others began to apply recombinant DNA technology to the problem. Today the work on HBV stands as the most successful application of recombinant DNA technology to medical virology.

PIERRE TIOLLAIS and MARIE-ANNICK BUENDIA are pioneers in the use of recombinant DNA technology to study hepatitis viruses. Tiollais is professor and director of an INSERM research unit at the Pasteur Institute and professor of biochemistry at the University of Paris 7. After receiving his medical degree from the University of Paris in 1968, he began studies of nucleic acid synthesis. He was the first to clone the human hepatitis B virus. Buendia, a director of research of the CNRS at the Pasteur Institute, received her doctorate in biochemistry from the University of Paris in 1977. She has played a particularly important role in illuminating how the woodchuck hepatitis virus may induce liver cancers.

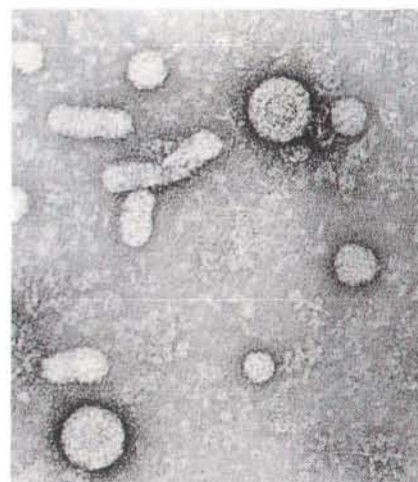
People infected with HBV are often unaware of it. After a two- to six-month incubation period, HBV infection can lead to acute hepatitis and liver damage, which cause abdominal pain, jaundice, elevated levels of certain enzymes in the blood and other symptoms. At this stage, HBV infection can be diagnosed by detecting HBsAg in a patient's blood serum. More frequently, the disease remains permanently asymptomatic.

In very rare cases, the HBV infection causes fulminant hepatitis—a rapidly progressive, often fatal form of the disease in which massive sections of the liver are destroyed. The liver damage is not indicative of a more virulent form of the virus but is instead the consequence of a stronger immune response by the host: white blood cells called "killer" T lymphocytes attack infected cells bearing the viral antigens.

Usually a patient with acute hepatitis will recover completely. The clinical and biological symptoms of the infection gradually disappear as more antibodies against the virus are produced. After recovery, the patient continues to produce low levels of those antibodies, which maintain immunity to HBV for several years. In the event of a new HBV attack, the antibody level rises quickly and neutralizes the virus.

In some patients, however, the high levels of viral antigen in the blood persist for several years or even for a lifetime, yet no antibodies against HBsAg ever appear. The virus survives in the liver, and the patient becomes a chronic carrier. The mechanism by which such a chronic state establishes itself is not well understood, but it seems to be related to a weak immune response. That explanation may account for why the infection becomes chronic in about 80 percent of infants, whose immune systems are immature, but in only 5 to 10 percent of adults.

Chronic infections can assume several different forms. Some chronic carriers are healthy: they experience limited liver damage and have no functional deficiencies. Others acquire chronic persistent hepatitis, which is general-



HEPATITIS B VIRUSES have a double-walled structure consisting of an outer envelope and an inner shell, or capsid, that contains the viral DNA. Many incomplete viral particles also appear in the blood of infected persons, as shown in the micrograph (*above*). These particles are small spheres or long filaments without an internal structure, and only two of the three characteristic viral envelope proteins appear on their surface.

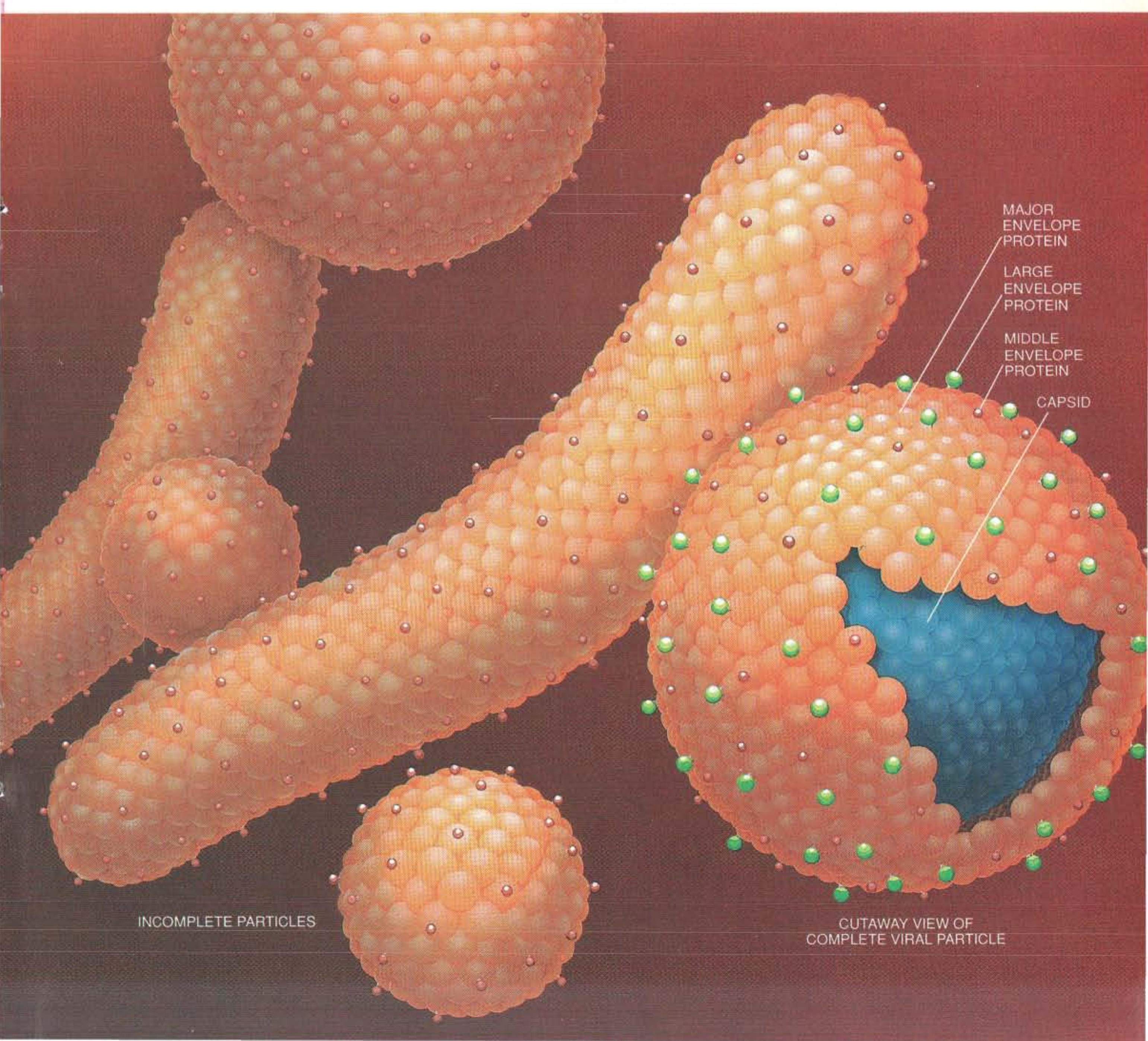
ly asymptomatic but sometimes causes fatigue. In the worst cases, chronic active hepatitis develops and can lead to cirrhosis of the liver and finally to hepatocellular carcinoma (HCC), a primary liver cancer. Usually the cancer does not develop until after a 30- to 50-year latency period, but a few cases of HCC have been observed in children.

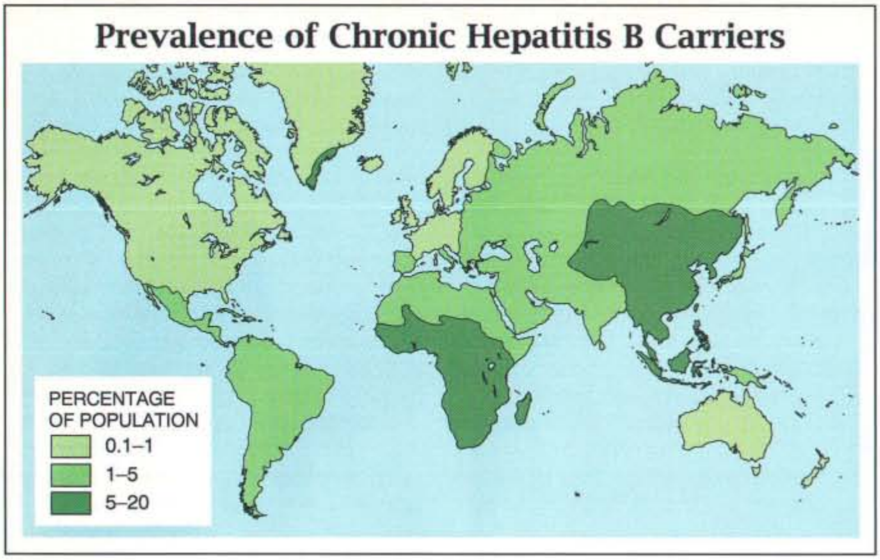
Today the link between chronic HBV infection and the development of HCC is very clear. The percentage of chronic carriers is higher among patients with HCC than among the rest of the population. Moreover, in an epidemiological study performed in Taiwan, R. Palmer Beasley of the University of Texas at Houston has shown that for chronic carriers the risk of acquiring HCC is 100 times higher than normal. Other studies have also found a consistent and specific causal association between HBV and HCC. HBV is therefore one of the few viruses known to cause a specific cancer in humans.

The gravity of the link between HBV and cancer becomes clear when the huge numbers of infected persons are considered. Worldwide, chronic HBV infections affect almost 300 million people, three quarters of whom are in Asia. The incidence of the condition varies greatly among different regions. In southeast Asia and tropical Africa, chronic carriers of the virus represent 10 percent or more of the population, whereas they make up less than 1 percent in North America and most of western Europe.

In the Third World the virus is most often transmitted from an infected mother to her infant in the baby's first month, primarily during birth. If an infant is female and infected, then she will probably become a chronic carrier and transmit HBV to her own offspring when she reaches childbearing age. Mother-to-child transmission is not the only mechanism: because the virus can be found in blood, saliva and semen, any intimate or sexual contact can pass on the disease. That fact explains the easy spread of hepatitis B within a family or small community.

In Western industrialized countries, other mechanisms of HBV transmission become much more important. High-risk populations consist of persons in direct contact with chronic car-





Chronic HBV carriers are most common in developing countries where these infections are endemic. About 225 million chronic carriers of HBV are in Asia alone.

riers or with their blood samples (such as nurses, physicians and dentists), the recipients of blood or blood products (such as hemophiliacs and patients who receive transfusions or dialysis treatments), intravenous drug abusers, homosexuals and persons with multiple sex partners. For the population outside those groups, the risk of contracting hepatitis is low. In fact, the epidemiology of hepatitis B is very similar to that of acquired immunodeficiency syndrome (AIDS)—which explains why hepatitis B infection is common among patients with AIDS or the AIDS-related complex. Hepatitis is far more contagious than AIDS, however.

Hepatitis B is therefore primarily a disease of infants in developing nations, whereas in Western countries it is mostly confined to adults. This distinction has practical ramifications for future vaccination strategies. In the Third World, mass vaccinations would be necessary, but in Western countries only the high-risk population needs to be protected.

Antiviral vaccines work by sensitizing a person's immune system against viral molecules. Much of the early work on HBV therefore concentrated on deciphering the structure and life cycle of the virus to identify good vaccine targets.

All viruses are cellular parasites made up of an RNA or DNA chromosome (genome) packed in a protein coat. To proliferate, a virus must penetrate the cell and use the cellular machinery to synthesize the proteins of the coat and the viral genome. After a genome is packaged inside its coat, the new viral particles leave the cell and can infect neighboring cells.

HBV has a double-walled structure consisting of two concentric protein coats. The envelope, or outer coat, contains three proteins that are designated major, middle and large. HBsAg, the surface antigen, is found on all three of the proteins. The capsid, or inner coat, consists of a single core protein that surrounds and interacts with the viral DNA.

The HBV genome, which was first isolated in 1974 by William S. Robinson of the Stanford University School of Medicine, is a circular DNA molecule composed of only about 3,200 nucleotide subunits. It is the smallest-known animal virus genome—the genome of the common herpesvirus, for example, is 50 times larger. Like the DNA molecules in most organisms, the genome of HBV consists of two base-paired strands, but it has an unusual feature: one strand is longer than the other. The short "plus" strand, which can vary in length, is only 50 to 80 percent as long as its mate, the "minus" strand. (As we shall explain later, this unusual structure is a consequence of the unique replication mechanism used by the virus.) The circular structure of the genome is maintained by base pairing of the strands at one end.

Using recombinant DNA techniques, in the late 1970s our research group succeeded in cloning, or copying, the HBV genome in *Escherichia coli* bacteria. This breakthrough made it possible to produce large quantities of the virus and its components for further study. Patrick Charnay of our laboratory, in collaboration with Francis Galibert of the St. Louis Hospital in Paris, subsequently determined the complete nucleotide sequence of the HBV genome, thereby providing the first available information about the genetic organization of the virus.

The HBV genome is a miracle of compactness. It consists of only four potential genes, termed *S*, *C*, *P* and *X*, which overlap extensively. The regulatory sequences that control the production of the viral proteins and the replication cycle are also embedded in these coding sequences.

Gene *S* encodes, or carries the blueprints for, the major envelope protein and includes all the specifications for HbsAg. A sequence of about 500 nucleotides preceding the *S* gene can also be transcribed along with it. As Wolfram H. Gerlich of the University of Göttingen has shown, this "upstream" sequence can be divided into two regions, pre-*S1* and pre-*S2*, that are involved in the synthesis of the other envelope proteins: the middle protein is encoded by pre-*S2* and the *S* gene, and the large protein is encoded by pre-*S1*, pre-*S2* and the *S* gene. The pre-*S1* region also plays an important role in the entry of the virus into a liver cell, as shown by A. Robert Neurath of the New York Blood Center.

Gene *C* encodes the protein of the capsid. Like *S*, the *C* gene is preceded by a short pre-*C* region that encodes a hydrophobic peptide involved with the assembly of the viral particle. Gene *P*, which is so long that it includes parts of all the other genes, encodes enzymes essential to the viral replicative cycle. Gene *X* covers the cohesive ends of the viral DNA strands. Its protein product stimulates the expression of all the viral genes by interacting with a specific DNA sequence found in the the HBV genome.

In the life cycle of HBV the synthesis of viral proteins is tightly regulated at the transcriptional and translational level. Two types of messenger RNAs (mRNAs) copied from the viral genome are known. The smaller one, about 2,100 nucleotides long, encodes the major and the middle envelope proteins. The larger mRNA, which is about 3,500 nucleotides long, is surprising: it is actually longer than the complete genome and contains a terminal repeat of about 100 nucleotides. The large mRNA encodes the capsid protein and the products of gene *P*, as shown by Harold E. Varmus of the University of California at San Francisco and Heinz Schaller of the University of Heidelberg. It also represents an intermediate in the replication of the viral DNA, as we shall soon describe.

Transcriptional "enhancer" elements in the HBV genome activate the expression of all viral genes and operate preferentially in liver cells. Other regulatory elements modulate the levels of individual proteins. A good illustration of this complex mechanism is provided by the variations observed in the amounts of large, middle and major proteins of the envelope made in infected liver cells. The large proteins are produced in small quantities and are found only on the surface of complete, infectious viral particles. In contrast, the major proteins (and, to a lesser extent, the middle proteins) are made in great excess and secreted in smaller (22-nanometer) particles, which are more abundant than complete particles in the serum.

One important question we sought to answer was how the various HBV genes are expressed. We have found that transgenic mice—animals that have had part or all of the HBV genome added to their own DNA—are powerful tools for studying the expression of the viral genes in normal host tissues.

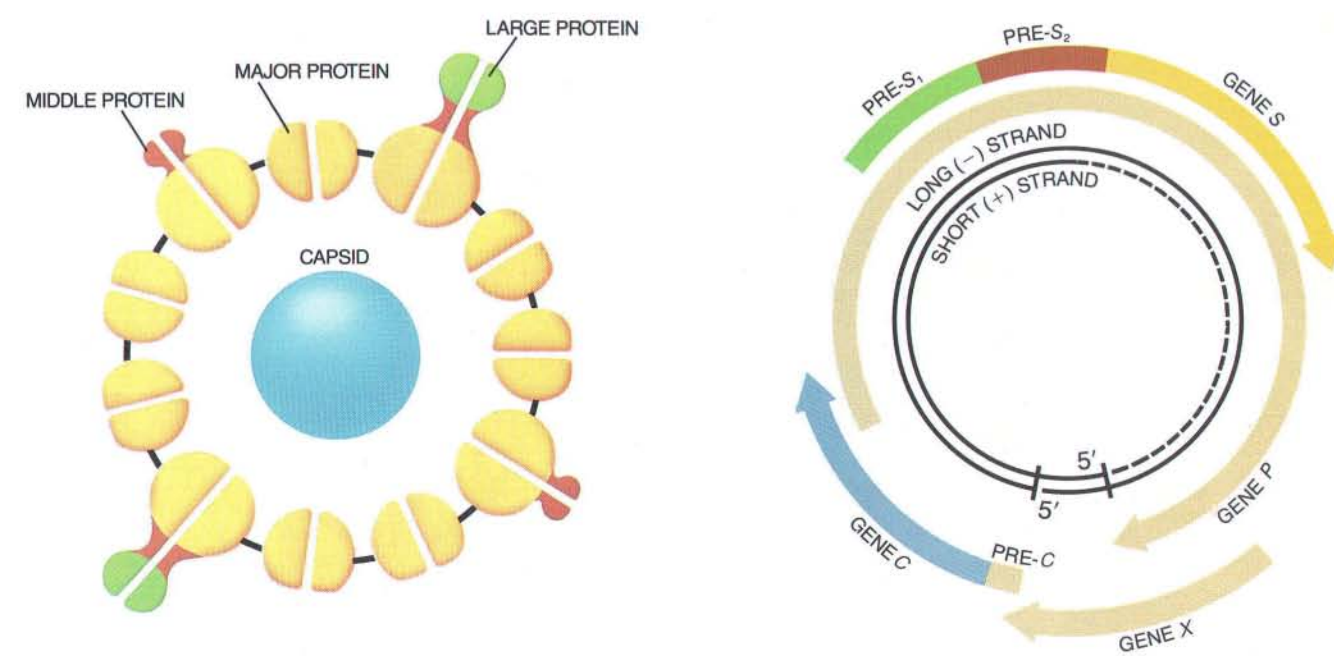
Working with transgenic mice, Christine Pourcel of our laboratory has studied why HBV preferentially infects certain hosts and host tissues. She has shown that gene *S* is expressed at very high levels only in liver tissue and under the control of steroid hormones. These discoveries clarify studies of human HBV carriers, which have shown that men (whose natural steroid hormone levels are higher than women's) run a greater risk for chronic HBV infection, liver damage and HCC.

In other strains of transgenic mice, viral replicative forms were observed by Pourcel in kidney and heart cells as well as in the liver. These results, too, mirror the affinity of HBV for different tissues in humans: although HBV is most abundant in the liver, the viral DNA and proteins have also been found in the kidney, spleen, pancreas, skin, bone marrow and circulating blood cells. The blood cells might be the first targets of HBV infection; viruses persisting in these sites might also be able to reemerge and attack the new livers of hepatitis patients who have received transplants. The infection of certain white blood cells may be directly involved in some other diseases, such as aplastic anemia and polyarthritis nodosa, and in the development of AIDS and AIDS-related complex.

Some understanding of HBV has come from studies of similar viruses, called hepadnaviruses, that cause hepatitislike conditions in animals. Hepadnaviruses that infect the woodchuck, the ground squirrel, the Peking duck, the tree squirrel and the heron have been discovered in recent years. The viral particles of these animal viruses strongly resemble HBV particles. Like HBV, they have genomes that are circular, partially single-strand DNA molecules, and almost all of them have identical genetic organizations.

Hepadnaviruses reproduce by an extraordinary mechanism. Most DNA viruses directly copy their genomes with enzymes called polymerases: the polymerases use the original DNA strands as templates for assembling new complementary strands. Yet hepadnaviruses use an indirect method, involving a strand of RNA as an intermediate. The mechanism was first identified in 1982 by Jesse W. Summers and William S. Mason of the Fox Chase Cancer Center in Philadelphia, who were working with the Peking duck virus.

The mechanism works as follows: After a hepadnavirus genome penetrates a cell, it migrates to the nucleus, where cellular polymerases transcribe it into a long RNA molecule. This 3,500-nucleotide RNA molecule is called the pregenome. The pregenome and a viral DNA polymerase are packed into a newly formed capsid and pass back into the cell's cytoplasm. There the polymerase goes to work, reverse transcribing the pregenome into a new minus DNA strand. As soon as the minus strand forms, the pregenome is destroyed by enzymes. DNA polymerases then begin to reconstruct the plus strand, using the minus strand as a template. The capsid and the viral DNA are finally enclosed in a new outer envelope as the virus leaves the liver cell. When the virus exits, however, the elongation of the plus strand stops immediately. Conse-



VIRAL PROTEINS in the hepatitis B infectious particle (*left*) are encoded by overlapping genes in the viral DNA (*right*). The color code shown here matches proteins to their genes. Other genes make products essential to the viral life cycle.

How the Hepatitis B Virus Replicates

1 The hepatitis B virus infects a liver cell.

2 Enzymes extend the short DNA strand of the viral genome.

3 The viral DNA migrates to the cell nucleus, where it is copied into RNA.The 3.5-kilobase strand of RNA becomes the pregenome intermediate for further replication.

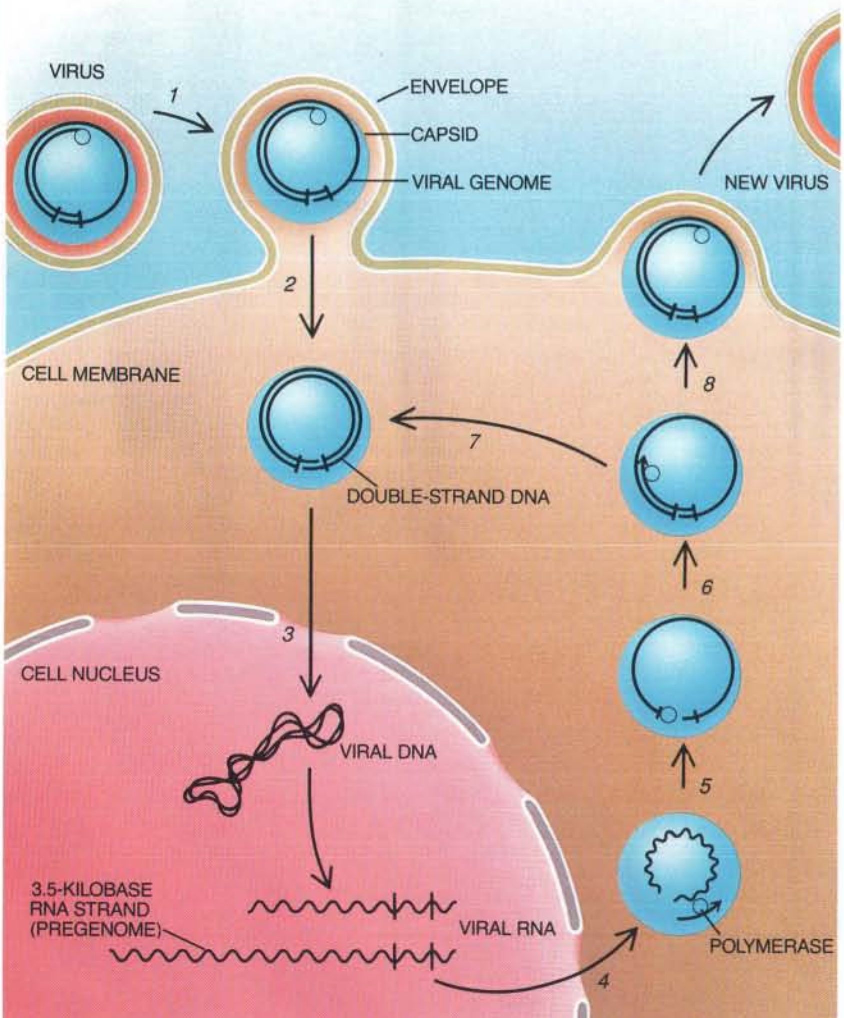
4 The pregenome is packaged in a freshly made capsid. Polymerase enzymes begin to make a DNA copy of the pregenome.

5 The new DNA strand is a duplicate of the original long strand in the viral genome. The pregenome disintegrates as the new DNA is completed.

6 Polymerases begin reconstructing a complementary DNA strand from the long-strand template.

7 The viral DNA may persist in the cell long enough to become fully double stranded. It can then return to the nucleus for another round of replication.

8 If the new viral particle exits the cell instead of replicating again, the capsid is enclosed in a fresh envelope. The extension of the short DNA strand stops immediately.



quently, the length of the plus strand is variable.

In its broadest outline, the amazing replication mechanism of the hepadnaviruses is the mirror image of the one used by the retroviruses, such as the AIDS virus, which have an RNA genome and use a DNA molecule as an intermediate. Hepadnaviruses and retroviruses have several other analogous features as well. Both types can infect cells chronically without destroying them. The order and function of the retroviral genes *gag*, *pol* and *env* are directly comparable to those of the hepadnaviral genes *C*, *P* and *S*. (*Gag* encodes the retroviral capsid protein, *pol* encodes a transcriptional enzyme and *env* encodes the envelope proteins.) Both families of viruses can also cause some cancers.

Epidemiological studies clearly indicate that chronic infection with HBV or the similar hepadnaviruses is sufficient to induce liver malignancies. Mammals chronically infected with hepadnaviruses, for example, show a high incidence of liver tumors: more than 80 percent of infected woodchucks develop primary liver cancers within two years. Hans Popper of Cornell University, John L. Gerin of Georgetown University and their colleagues have also shown that HCC could be experimentally induced in woodchucks through inoculation with the hepatitis virus for that species.

Those experiments confirmed the cancer-causing properties of a hepatitis virus and ruled out the need for carcinogenic cofactors. In the human disease, other influences may also be at work: the period of latency preceding the development of HCC is much longer in people than in woodchucks, and the tumors frequently appear in cirrhotic livers. Environmental carcinogens and excessive consumption of alcohol may still be involved in the genesis of human HCC.

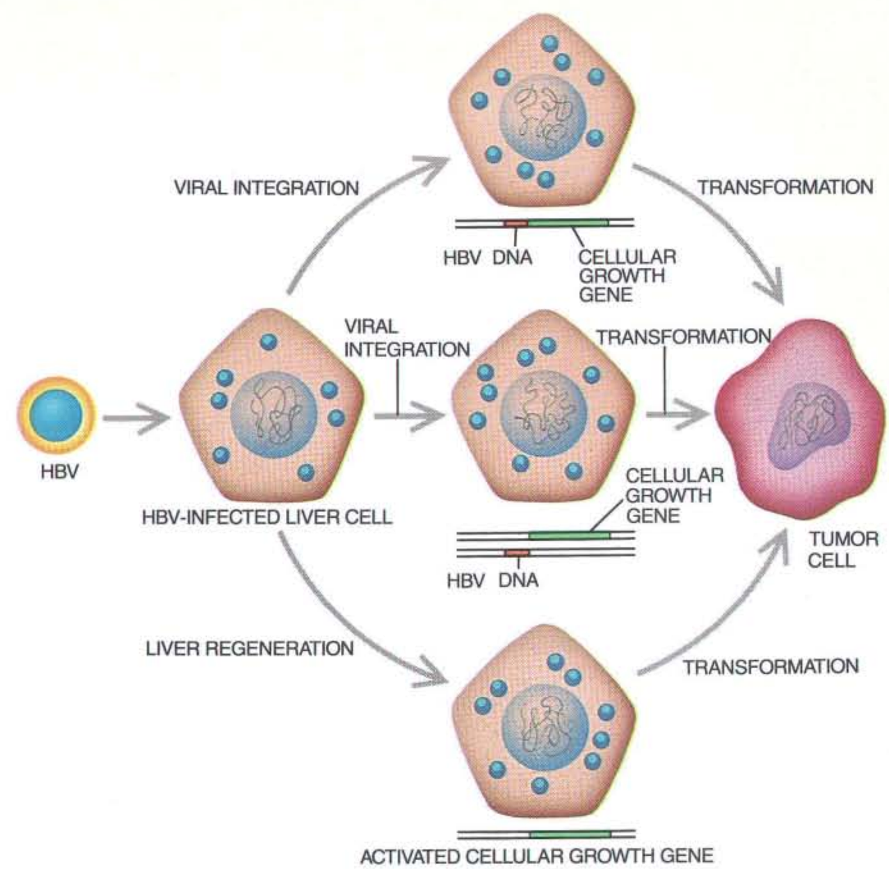
The question of how HBV induces cancer is still open to debate. HBV might directly trigger tumor development. Alternatively, the liver tumor may arise indirectly because of the chronic inflammation, cirrhosis and cell regeneration taking place in the diseased tissues.

Many tumor viruses carry oncogenes, genes that can directly transform infected body cells. (Similar oncogenes are also found naturally in cells, where they seem to control growth and development.) The HBV genome, however, does not apparently carry an oncogene. Moreover, the long latency period between an HBV infection and the appearance of a tumor is not compatible with the idea that a viral oncogene

causes HCC. It is nonetheless conceivable that the gene-activating protein produced by HBV gene *X* might be involved at some early stage of the tumorigenic process.

A virus can also transform a host cell by introducing its DNA into the cellular chromosomes. The insertion of viral sequences adjacent to cellular oncogenes can deregulate their expression and lead to uncontrolled growth. That strategy is typical of some retroviruses that induce leukemias and carcinomas in mammals and birds after a latency period. Although HBV does not normally need to integrate genetic material into its host's genome to complete its life cycle, such integrated forms might be incidental by-products.

Christian Brechot, while working in our laboratory, David A. Shafritz of the Albert Einstein College of Medicine and William J. Rutter of the University of California at San Francisco have analyzed DNA molecules from HCCs and found that the viral sequences were integrated into the chromosomes of the tumor cells. Although such sequences were mainly observed in tumors from chronic HBsAg carriers, they were also found in tumors from HBsAg-negative patients. Integrated viral DNA was also occasionally detected in the livers of chronic HBV carriers, which suggests that the DNA integration may occur before or at a very early stage of tumor formation.



THREE EXPLANATIONS for how hepatitis B infections could transform liver cells into tumor cells have been proposed. Viral DNA may turn on cellular growth genes by integrating itself adjacent to those genes (*top pathway*) or elsewhere in the genome (*middle pathway*). Another possibility is that the cells regenerating after the infection may erroneously activate their own growth genes (*bottom pathway*).

Analyses of tumors and cell lines have shown that HBV can insert its DNA at many different sites in human chromosomes. It often induces substantial genetic rearrangements, such as chromosomal deletions, translocations or amplifications—abnormalities common in human tumors and cancers. The consequences of those genetic events are still unknown.

In one early human HCC, Anne Dejean of our laboratory found a single insertion of HBV DNA into one member of a gene family that makes receptors for thyroid and steroid hormones. She was able to isolate and characterize that gene and identify its specific product as a receptor for retinoic acid.

Retinoic acid exerts a powerful influence on cell differentiation and proliferation, and excesses of it during early pregnancy have been linked to birth defects. It can also make many cultured, abnormal cells return to their normal form and so is used to cure some leukemias. A mutation in a retinoic acid receptor gene caused by the insertion of HBV DNA could play a part in the genesis of human HCC.

Brechot, who is now at the Necker Hospital in Paris, observed the similar integration of HBV DNA into the cyclin A gene in another human HCC. Cyclin A is involved in cellular growth, and its abnormal expression can deregulate the normal control of cell proliferation. Much more information about the frequency of such viral insertions into cellular genomes is needed.

We have obtained further clues to the role of viral integration in HCC from a study of woodchuck tumors induced by woodchuck hepatitis virus. Viral DNA invariably integrates itself in these tumors, and in about 30 percent of the cases, we have identified the target sites for viral integration as two members of the cellular *myc* oncogene family. Those genes are normally involved in the control of cell growth and differentiation, but they can also contribute to the formation of lymphoid tumors and various carcinomas.

A part of the viral genome harboring the enhancer element inserts itself near the *myc* oncogene without modifying the gene's code. The synthesis of normal Myc proteins, now triggered by viral information, escapes its normal cellular control. Consequently, the cell overproduces the Myc proteins that induce unregulated cell growth. This mechanism is basically identical to the way in which retroviruses are known to cause lymphoid tumors in rodents and birds.

There is as yet no evidence that a similar pathway leads to the development of human HCC: extensive studies have failed to reveal insertional mutations in *myc* oncogenes caused by HBV DNA. Nevertheless, other human oncogenes might be activated either directly or indirectly by the HBV insertion. Peter Hans Hofschneider of the Max Planck Institute in Munich has shown that integrated HBV DNA can produce modified proteins that stimulate the expression of both viral and cellular genes, including potent oncogenes. A more indirect role for the virus was reported by Francis V. Chisari of the Research Institute of Scripps Clinic: in transgenic mice, overproduction of the HBV large protein leads to chronic liver cell injury and finally to HCC.

In DNA transfer experiments, Kenichi Matsubara of Osaka University has

demonstrated the activation of two other cellular oncogenes: *lca*, a gene that has so far been implicated only in HCC, and *hst-1*, which is also involved in the formation of stomach cancers. Other genetic alterations that are not obviously linked with HBV infection have also been observed in human HCC tumors. Further studies are needed to understand the viral role in carcinogenesis at the molecular level.

The preceding experiments demonstrate how instrumental recombinant DNA technology has been for understanding the basic biology of HBV. The technology also has practical applications in the development of new diagnostic techniques and vaccines.

Diagnostic tests for hepatitis B have been improved by the technique called molecular hybridization, which capitalizes on the high affinity that short strands of cloned DNA have for related complementary DNA. Cloned, labeled viral DNA can serve as a sensitive, reliable probe for detecting infectious viral particles in blood serum. Currently molecular hybridization using HBV DNA is routinely used to identify recent hepatitis B infections. Because the identification of viral DNA also allows an accurate estimation of how much viral replication is taking place, molecular hybridization is also suitable for following the progress of antiviral treatments.

The sensitivity of the method can be further improved by using the polymerase chain reaction technique, which is a simple enzymatic procedure for duplicating even traces of specific DNA sequences at exponential rates [see "The Unusual Origin of the Polymerase Chain Reaction," by Kary B. Mullis; SCIENTIFIC AMERICAN, April 1990]. With this technique, hybridization-based diagnostic tests can increase their sensitivity more than 1,000-fold: they can detect as few as 100 viral particles in a milliliter of serum, which is considered to be the minimum infectious concentration.

Recombinant DNA technology is also developing new vaccines. Such progress began in 1970, when Saul Krugman and his co-workers of New York University demonstrated that the heat-inactivated serum of a chronic carrier could protect susceptible persons against HBV. Later, several groups showed the protective effect of purified viral particles carrying HBsAg in chimpanzees.

In 1976 Philippe Maupas and his co-workers of the University of Tours reported the first results of vaccination in humans. Their vaccine consisted of defective viral particles—essentially, empty envelopes bearing HBsAg—purified from the blood of chronic carriers and treated to eliminate all infectious particles. It was the only antiviral vaccine ever developed from the blood of patients. Maupas and his colleagues later clearly demonstrated the safety and efficacy of the vaccine for a very large group of recipients, including many infants in regions where HBV is endemic and high-risk populations where it is not endemic. (The vaccine was less effective in patients on dialysis, probably because of their weakened immune systems.)

Although the serum-derived vaccine is effective, producing it poses several problems: the supply of human serum from chronic carriers is limited, and the purification procedure is long and expensive. Moreover, each batch of vaccine prepared from a different stock of infected human serum must be tested on chimpanzees to ensure its safety. For all these reasons, the idea of applying genetic engineering to produce a hepatitis B vaccine was a very attractive one to researchers.

Several genetic engineering techniques can be used to develop vaccines, but the choices are more restricted in practice because HBsAg is not fully immunogenic unless it retains its complete natural structure. Genetically engineered bacteria can mass-produce HBsAg, but the antigen does not have its correct chemical form or structure. Yeast and mammalian cells, however, can produce useful HBsAg proteins.

HBV vaccines produced in yeast or mammalian cells are now on the market. Rutter obtained a vaccine by producing one in *Saccharomyces cerevisiae* (baker's yeast). In 1984 Marie Louise Michel of our laboratory genetically constructed a line of Chinese hamster ovary cells that produce HBV antigens. The recombinant HBV DNA that she inserted contained both the *S* gene and the pre-*S2* region. The resulting particles therefore carry both the major and the middle envelope proteins and bear both the HBsAg and pre-*S2* antigenic determinants.

That fact may be significant because David R. Milich of the Scripps Clinic has shown in mice that the pre-*S2* determinant elicits a stronger immune response than HBsAg. Moreover, in mouse strains that do not produce antibodies against HBsAg, immunization with Michel's recombinant particles can circumvent the nonresponsiveness. This property could be important for the vaccination of people who are unresponsive to that antigen.

Another intriguing recombinant DNA approach would be to construct a live vaccine using the vaccinia virus or the adenovirus as vectors. Genetic engineering could make these harmless viruses express HBV antigens. This technique has not been applied to vaccines for humans until now. It should also be possible to chemically synthesize amino acid sequences that would duplicate the antigenic determinants of HBsAg, and these could provide an inexpensive HBV vaccine. Unfortunately, such an approach is not practical today, because the immunogenicity of such artificial peptides, like those made by bacteria, is extremely low.

In addition to producing vaccines that can prevent HBV infection, recombinant DNA technology may also be on the verge of finding a cure for chronic carriers. In August 1990 Robert P. Perillo of Washington University and his colleagues announced that daily treatments with alpha interferon, a genetically engineered protein that boosts immune responses, had eliminated or inhibited infection in more than a third of 85 chronic HBV carriers. Further research on the use of interferon against HBV promises to refine and improve such treatments.

During the past decade, the evolving understanding of the molecular biology of HBV has found medical applications, particularly for preventing the infection. Many of the benefits have barely been tapped. In developing countries, mass vaccinations with the recombinant vaccine—the first vaccine for human use ever developed by recombinant DNA technology—will go far in controlling hepatitis B. Such a campaign would bring twofold relief: the vaccinations would prevent not only the acute liver illness but also the associated cancer. We can hope that it would usher in an era of better medical care based on prudent applications of genetic engineering.

FURTHER READING

ADVANCES IN HEPATITIS RESEARCH. Edited by Francis V. Chisari. Masson Publishing USA, 1984.

HEPATITIS B. Edited by Robert J. Gerety. Academic Press, 1985.

THE HEPATITIS B VIRUS. Pierre Tiollais, Christine Pourcel and Anne Dejean in *Nature*, Vol. 317, No. 6037, pages 489–495; October 10, 1985.

THE MOLECULAR BIOLOGY OF THE HEPATITIS B VIRUSES. Don Ganem and H. E. Varmus in *Annual Review of Biochemistry*, Vol. 56, pages 651–693; 1987.

VIRAL HEPATITIS AND LIVER DISEASE. Edited by Arie J. Zuckerman. Alan R. Liss, Inc., 1988.

No other system of keeping up can compare with ours.

YOUR SYSTEM: a time-consuming, futile struggle to keep up with the information explosion.

The classic texts are convenient references—but the information they contain is obsolete before publication.

Like many physicians, you probably rely on the texts you first used in medical school. But even using the most recent editions, you find material that no longer reflects current clinical thinking—and they lack the latest information on such topics as herpes, oncogenes, AIDS, and photon imaging.

Reading stacks of journals alerts you to recent developments—but can't give you quick answers on patient management.

Struggling through the hundreds of journal pages published each month—even on only the really significant advances in the field—is arduous and memory-taxing. And it's a task that costs physicians valuable time—their most precious resource.

Review courses cover clinical advances—but, months later, do you recall the details of a new procedure or unfamiliar drug?

Seminars can also be costly and make you lose valuable time away from your practice—expenses that may amount to several thousand dollars. And, the speaker's skill often determines how much you learn.

OUR SYSTEM: a rewarding, efficient way to keep yourself up-to-date—and save hundreds of hours of your time for patient care.

A comprehensive, 2,300-page text in three loose-leaf volumes, incorporating the latest advances in medical practice as of the month you subscribe.

This superbly designed, heavily illustrated resource, called "the best written of all [the internal medicine] books" by *JAMA* (251:807, 1984), provides a practical, comprehensive description of patient care in 15 subspecialties. And, because the text is updated each month, the clinical recommendations reflect all the current findings. A practice-oriented index and bibliography of recent articles further enhance the efficiency of the text.

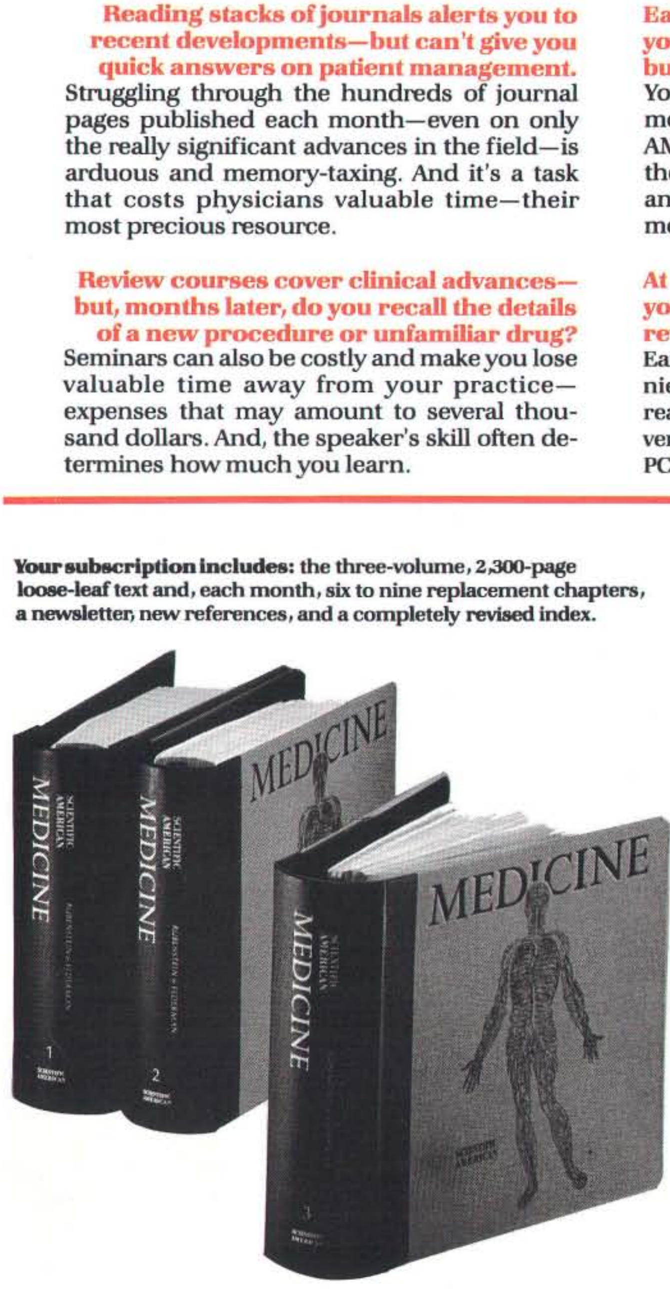
Each month, six to nine replacement chapters to update your text *plus* new references, a four- to five-page news bulletin, and a completely new index.

You'd have to read hundreds of journal pages each month—and memorize the contents—to get the same information SCIENTIFIC AMERICAN *Medicine* contains. With our updated text, you read only the information you really need. Our authors, largely from Harvard and Stanford, sort through the literature and monitor developments, incorporating the significant advances into our chapters.

At no additional cost, a 32-credit CME program, to save you valuable patient-care time and the expense of attending review courses.

Earn 32 Category 1 or prescribed credits per year with our convenient self-study patient management problems; each simulates a real-life clinical situation. Choose either the complimentary printed version or, at a modest extra charge, the disk version for your IBM® PC or PS/2, Macintosh™, or Apple® (or compatible).

Your subscription includes: the three-volume, 2,300-page loose-leaf text and, each month, six to nine replacement chapters, a newsletter, new references, and a completely revised index.



☐ Yes, I'd like to try the SCIENTIFIC AMERICAN *Medicine* system. Please enter my subscription at a first-year price of US$278* plus $7 shipping for a total of US$285.

☐ Also enroll me in the CME program. I'd prefer:
- ☐ the printed version at no additional charge.
- ☐ the disk version at US$97* additional. Computer type or compatible:
  - ☐ IBM® ☐ 5¼" ☐ 3½"
  - ☐ Macintosh™ (512K) ☐ Apple® IIc/IIe/IIGS
  - ☐ Apple® II

☐ Enroll me in the disk CME only at US$152.* (Note model above.)

☐ Check enclosed.* ☐ Bill me ☐ VISA ☐ MasterCard

Exp. Date______ Acct. No.______

Signature______ Specialty______

Name______

Address______

City______ State______ Zip Code______

*Add sales tax for DC, IA, IL, MA, MI or NY. Allow 6 weeks for delivery. Canada: add US$10 shipping, payment in U.S. funds or equivalent. Foreign: add US$22 shipping, U.S. funds only. All orders must be prepaid. Apple®, IBM®, and Macintosh™ are registered trademarks of their respective companies.

Or, call toll free 1 800 545-0554

SCIENTIFIC AMERICAN MEDICINE

415 Madison Avenue, New York, NY 10017

9117

Photochromic and Photosensitive Glass

Unlike ordinary glass, these glasses respond to light, resulting in structures that find uses in electronics, optics and the decorative arts

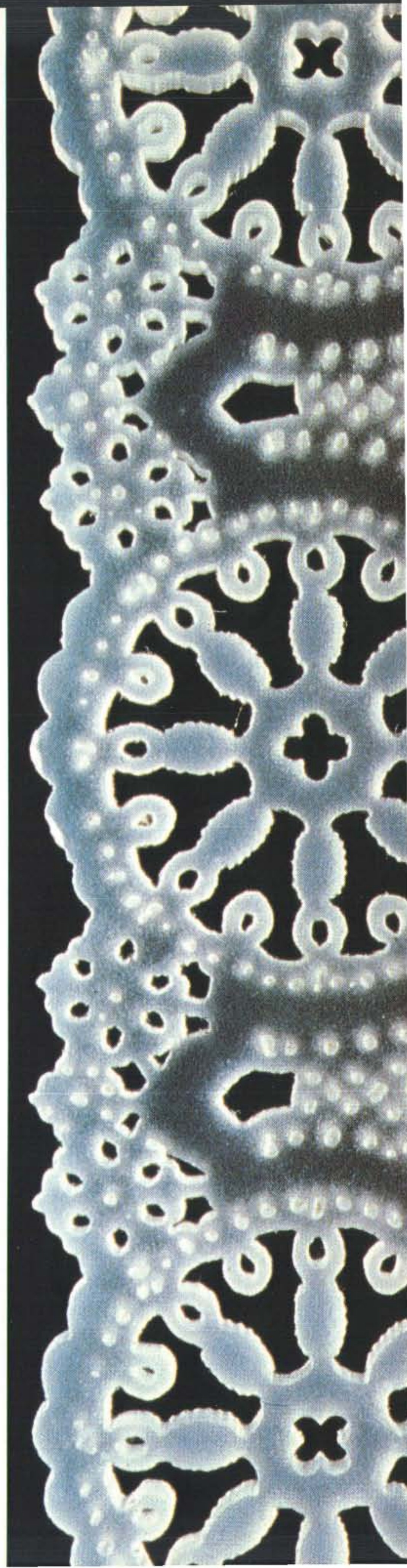
by Donald M. Trotter, Jr.

Since 2000 B.C. virtually every civilization that discovered the art of glassmaking exploited the many features of glass to make vessels, ornaments and sculpture. The distinctive characteristic that separates glass from clay is, of course, its transparency. Light passes through glass virtually unchanged; conversely, glass remains unaffected by the passage of light. But there are a host of less familiar glasses that do in fact respond to light and can undergo useful changes when illuminated. Such changes enable manufacturers to produce a variety of devices, from eyeglasses that darken outdoors to head pads in magnetic memory drives. Current research suggests such potential applications as self-shading windows for buildings, eyeglasses for some vision impairments and connectors for integrated circuits. Glass has found uses that could hardly have been conceived four millennia ago.

The glasses fall into two categories. Photochromic glasses temporarily change color when exposed to strong light. Photosensitive glasses, on the other hand, develop an invisible "latent image"; subsequent heat treatment develops this image, resulting in permanent structural or color changes—or both—in the glass.

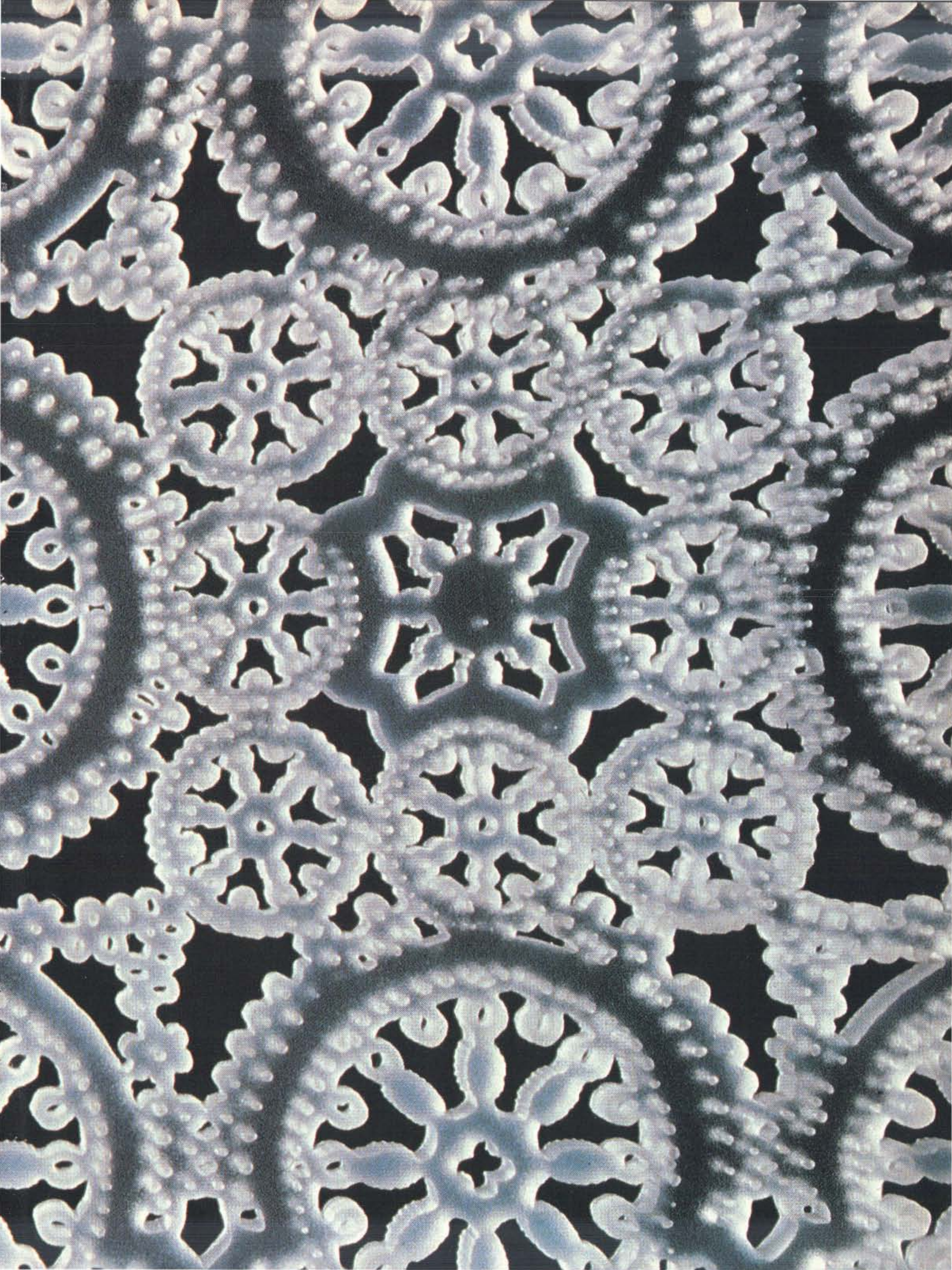
The chemical processes behind these glasses enable workers to custom-manufacture highly intricate parts for industrial applications with mass-production efficiency. Because the glasses are chemically machined, tolerances exceed those of parts mechanically cut and drilled, and holes can be made so small that a human hair cannot penetrate them. In addition, photochromic and photosensitive glasses are stable, inert, impermeable, strong and nonconducting, offering structural advantages over plastic and metal.

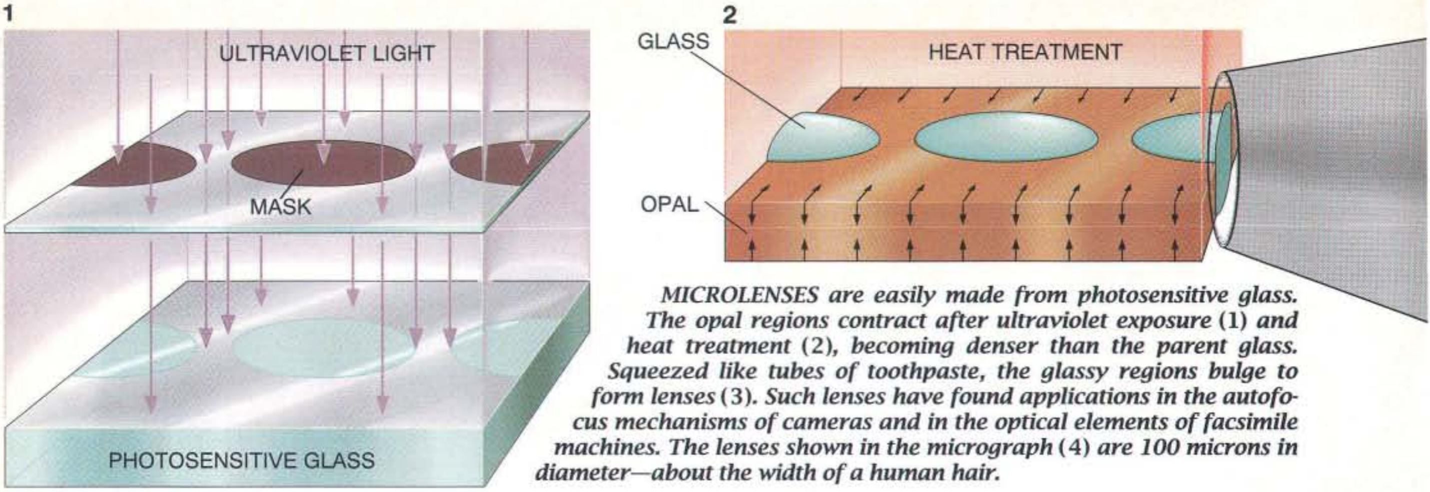
Scientists and engineers are currently experimenting with the properties of the glasses, searching for ways to refine their characteristics and to develop commercial applications. These workers are building on the chemical foundations first laid down in the 1940s. In the early work, S. Donald Stookey and other scientists at Corning Glass Works (now Corning, Inc.) in Corning, N.Y., had taken ordinary glass and experi-



PHOTOSENSITIVE GLASS ORNAMENT, shown here at about four times its actual size, is formed when ultraviolet light bathes unmasked regions. After heat treatment, those areas become opal. They are then leached away with hydrofluoric acid. The acid also etches the glass, giving it a frosted appearance.

DONALD M. TROTTER, JR., is a senior research scientist at Corning, Inc. Besides working on photochromic and photosensitive glasses, he spent several years investigating experimental capacitors, describing them in the July 1988 issue of SCIENTIFIC AMERICAN. He received his Ph.D. from the University of Texas at Austin and held a postdoctoral appointment at Cornell University. A science-fiction enthusiast, he has published several short stories in that genre.





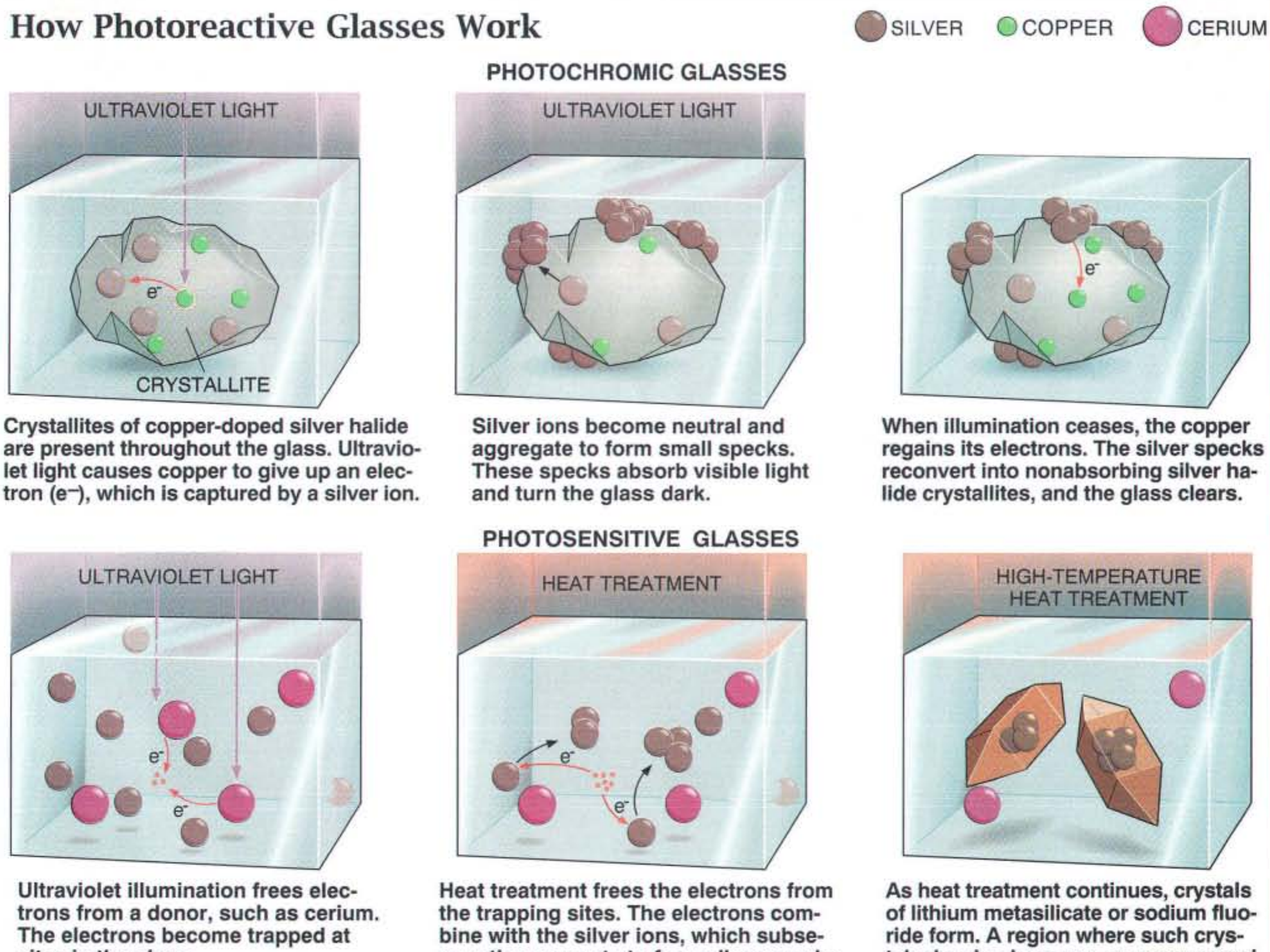
MICROLENSES are easily made from photosensitive glass. The opal regions contract after ultraviolet exposure* (1) *and heat treatment* (2)*, becoming denser than the parent glass. Squeezed like tubes of toothpaste, the glassy regions bulge to form lenses* (3)*. Such lenses have found applications in the autofocus mechanisms of cameras and in the optical elements of facsimile machines. The lenses shown in the micrograph* (4) *are 100 microns in diameter—about the width of a human hair.

mented with chemical additives. Stookey and his colleagues created families of glasses that react to light by changing color or forming a latent image. So successful was the basic chemical process worked out by Stookey that it remains in use today.

Like many common glasses, photochromic and photosensitive glasses consist predominantly of silicon dioxide—the primary constituent of ordinary sand. Silicon dioxide, however, is difficult to work. It melts at a very high temperature (above 1,600 degrees Celsius) and is extremely viscous. To reduce the melting temperature and create a more fluid melt, workers add alkali and alkaline earth oxides to the glass batch. Oxides of aluminum, boron, titanium and other metals are added to control such properties as refractive index, resistance to weathering and strength. Such oxides affect the ordering and the coordination of the molecules in the glass.

A small number of elements are mixed in to give this base glass its photochromic or photosensitive properties. The additives react or precipitate in response to illumination and heating. The base glass affects the properties of the resulting photochromic or photosensitive glasses by influencing the size and

How Photoreactive Glasses Work



Crystallites of copper-doped silver halide are present throughout the glass. Ultraviolet light causes copper to give up an electron (e^-), which is captured by a silver ion.

Silver ions become neutral and aggregate to form small specks. These specks absorb visible light and turn the glass dark.

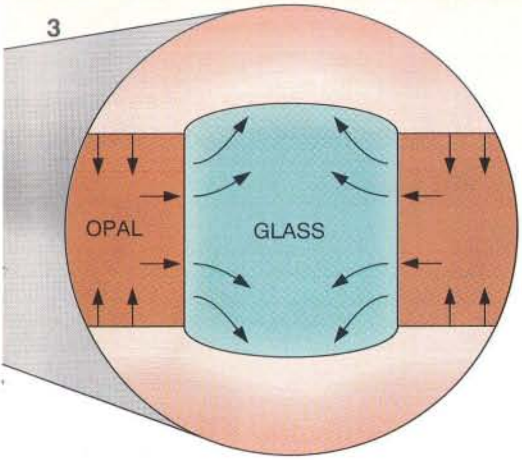
When illumination ceases, the copper regains its electrons. The silver specks reconvert into nonabsorbing silver halide crystallites, and the glass clears.

Ultraviolet illumination frees electrons from a donor, such as cerium. The electrons become trapped at sites in the glass.

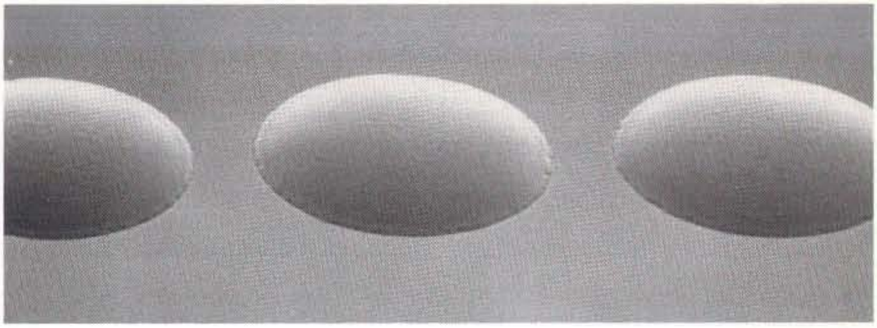
Heat treatment frees the electrons from the trapping sites. The electrons combine with the silver ions, which subsequently aggregate to form silver specks.

As heat treatment continues, crystals of lithium metasilicate or sodium fluoride form. A region where such crystals develop becomes an opaque opal.

3



4



composition of the precipitated particles or the rate of reaction.

To make a photochromic glass, one begins with a base-glass mixture that has a relatively large amount of boric oxide. Silver and copper in the form of nitrates or chlorides are added, as is a metal halide—a compound consisting of a metal and a halogen, such as fluorine, chlorine, bromine or iodine.

The batch is heated to a temperature of around 1,200 degrees C, at which point it melts into a homogeneous fluid whose consistency is that of thick syrup. Workers then pour the molten glass into molds, which press the liquid into various shapes—for example, slightly convex disks from which eyeglass lenses will later be ground. The mold chills the glass and causes it to harden and retain the shape.

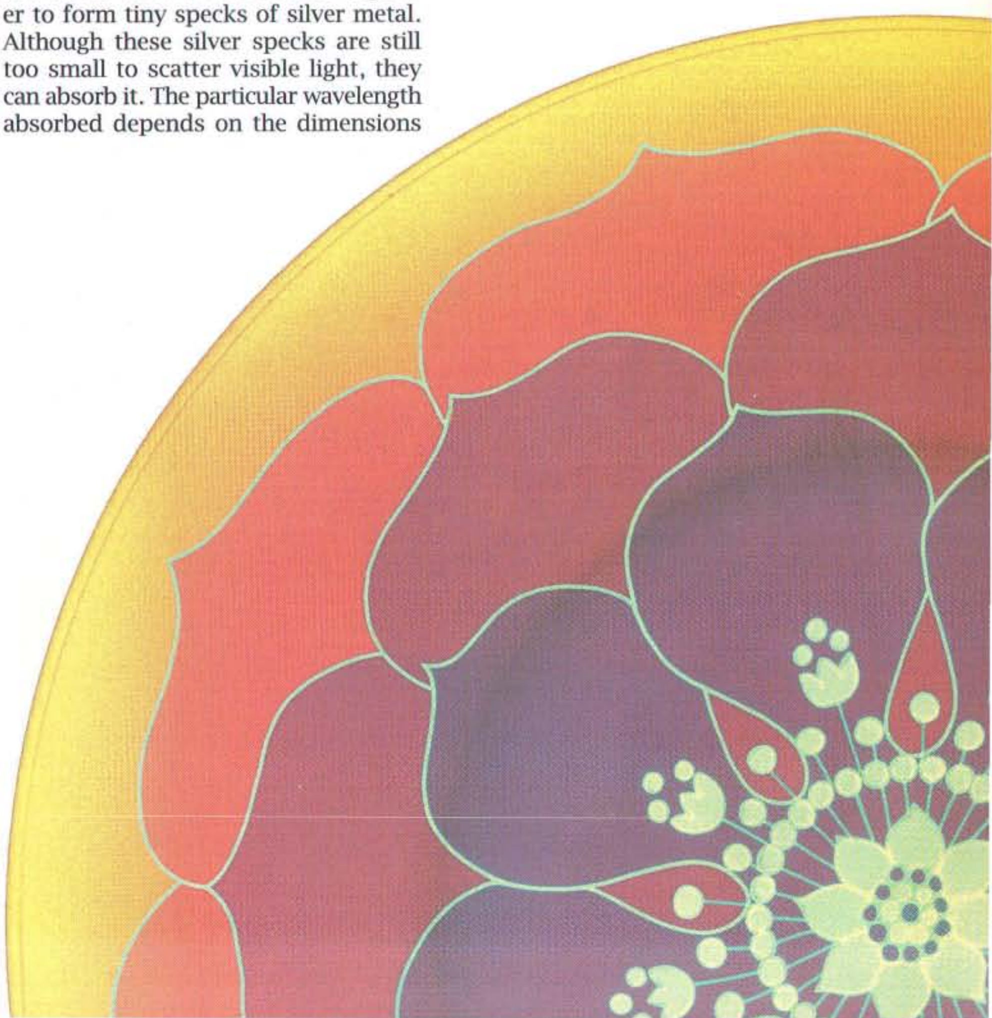
As the glass cools, boron changes the way it fits into the base-glass structure. Consequently, the halogens that were dissolved in the glass at the melting temperature become much less soluble. As a result, the halogens can come out of solution and react with the silver and copper. The reaction precipitates crystallites of silver halide that contain small amounts of copper halide.

Although the precipitation can, in principle, occur at room temperature, diffusion in the stiff glass matrix is so slow that many thousands of years would elapse before any crystallites formed. Workers are able to initiate rapid precipitation by reheating the glass to a temperature slightly below the softening point, typically around 600 degrees C, for about 30 minutes.

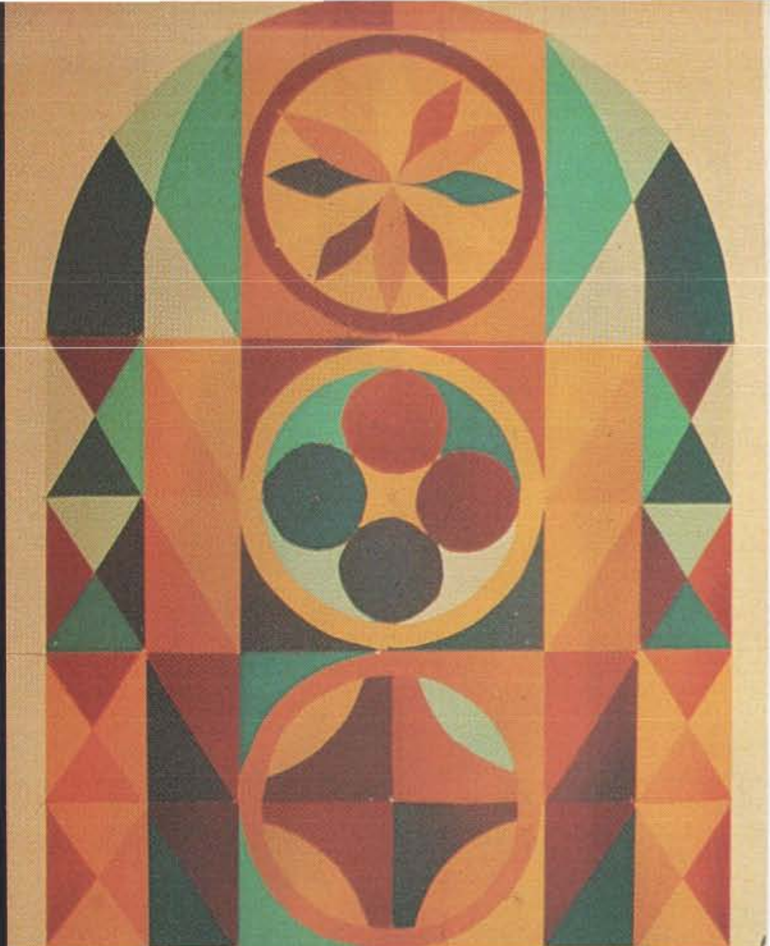
The crystallites, about 100 angstroms in diameter, are too small and too transparent to scatter or absorb visible light, which has a wavelength range of about 4,000 to 7,000 angstroms. The glass remains perfectly clear and colorless. The crystallites are not transparent to shorter wavelengths, however, and they absorb the ultraviolet light present in sunlight. When exposed to such light, some of the silver ions, positively charged and ionically bonded to negative halogen ions, gain an electron from a copper ion to become neutral atoms.

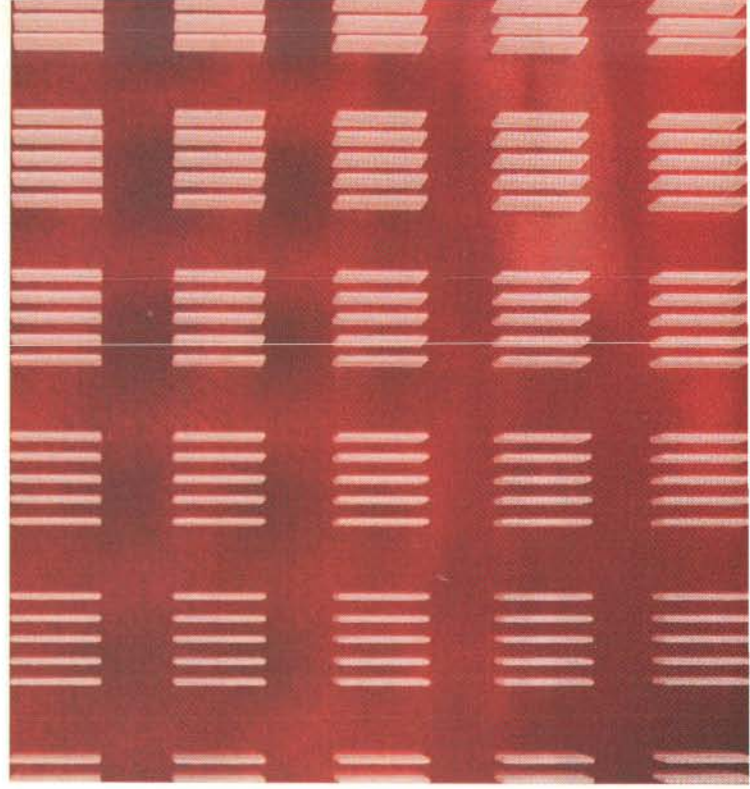
Hundreds to thousands of the neutral silver atoms then cluster together to form tiny specks of silver metal. Although these silver specks are still too small to scatter visible light, they can absorb it. The particular wavelength absorbed depends on the dimensions of the speck. Because many sizes and shapes are formed, all wavelengths are absorbed equally well. The glass consequently darkens to a gray color. When illumination ceases, the copper ions regain their lost electrons. The silver metal speck converts back to silver halide, and the glass quickly fades back to its original colorless state. Because nothing was lost from the crystallites during the process, the glass can darken and fade many thousands of times without degradation.

Improvements in the darkening and fading rates throughout the 1970s led to commercial applications, primarily in sunglasses. Recent work has focused on hydrogen firing of the glasses. Such firing produces permanent silver specks, which in turn give the glasses a deep color. The particular col-



DINNER PLATE was made from photosensitive glass. Exposure to ultraviolet light and heat has caused needlelike specks of silver to grow on precipitated sodium fluoride and sodium bromide crystals. The silver needles absorb a narrow region in the light spectrum and thus impart color to the glass. Variations in exposure affect the quantity of silver deposited and the needle size, so that different colors appear in the plate.





APPLICATIONS for photosensitive glasses vary widely. Because many colors are possible, such glasses are often used for dinnerware, holiday ornaments and art glass **(far left).** ***Integral louvers, about 3.2 millimeters wide, in glass doped with the additive sodium fluoride have been proposed for architectural uses, such as light control*** **(cen-**

or is determined by the duration and temperature of the firing. The lenses of some eyeglasses made in this way can minimize color distortion during the day but preserve vision at night. Others have spectral characteristics that seem to improve the comfort and visual acuity of patients with retinitis pigmentosa, a degenerative eye disease, as well as some other eye disorders that cause photophobia.

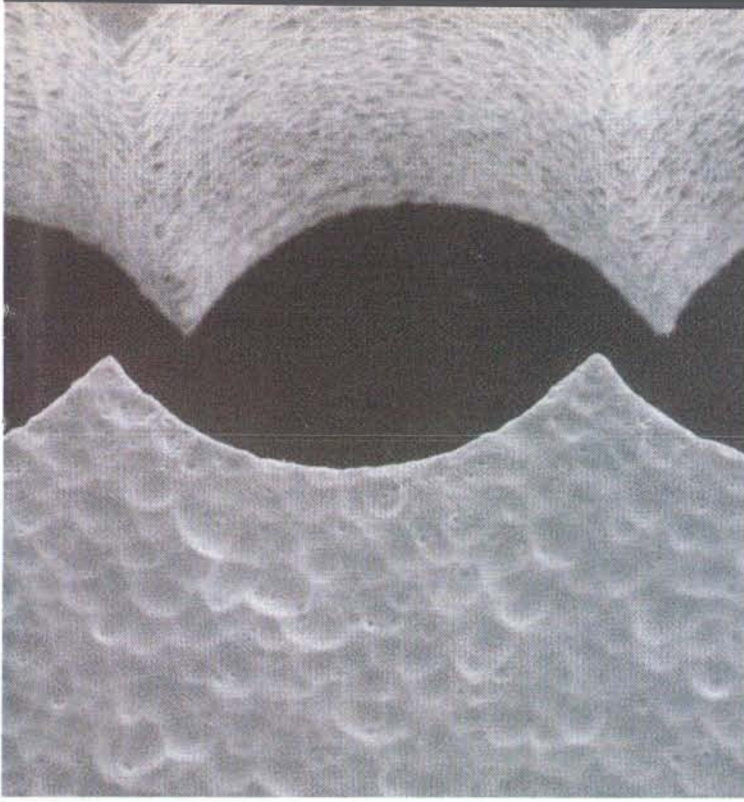
Researchers are also investigating the use of photochromic glasses in architecture and in the manufacture of automobile windows, although certain difficulties persist. For instance, the glasses tend to be sensitive to the ambient temperature. As a result, cold days will cause photochromic windows to darken more deeply than will warmer days, resulting in uneven lighting conditions inside.

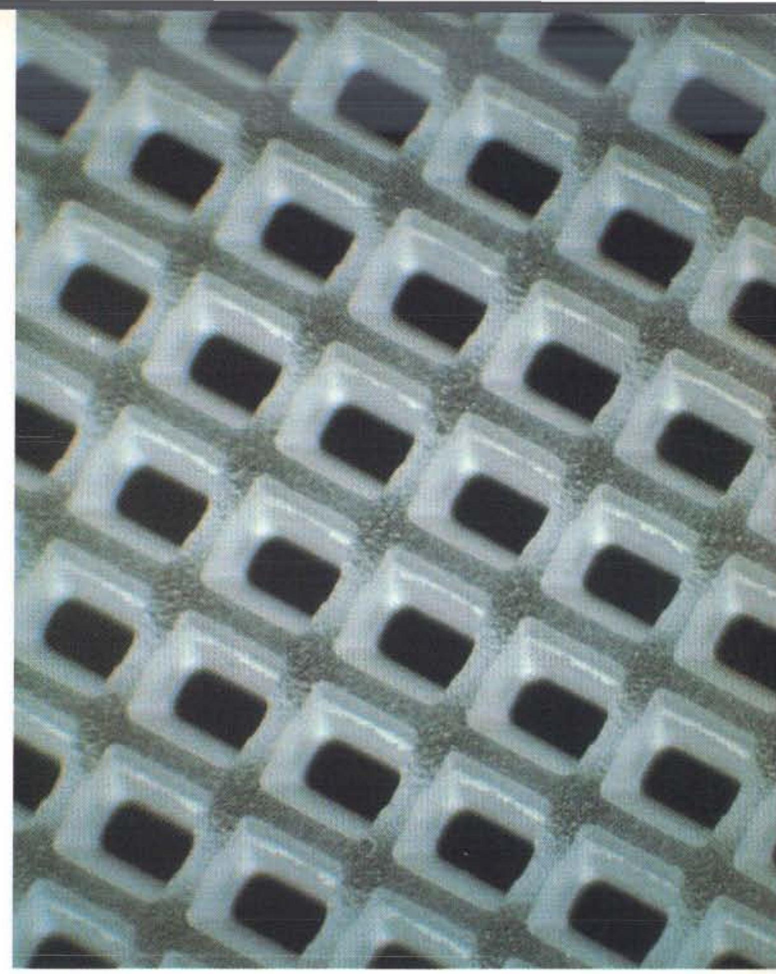
In contrast to photochromic glasses, photosensitive glasses change permanently. Silver ions again form the basis for the photochemical effects. But in photosensitive glasses, workers use cerium rather than copper as the electron donor for the silver, although theoretically several other elements can serve this purpose. When the glass is illuminated by ultraviolet light, some of the cerium ions give up an electron. The electrons released by the cerium are trapped at specific sites in the glass to form a latent image.

Researchers remain uncertain as to the nature of the trapping sites, which seem to be present throughout the glass. The sites also occur in photochromic glasses. But unlike those in photosensitive glasses, the silver ions in the crystallites of photochromic glasses are more efficient in gaining electrons than are the trapping sites.

Consequently, in photosensitive glasses, heat treatment is necessary to release the electrons from the traps; the electrons can then combine with the positive silver ions in the glass. Heating also enables the resulting neutral silver atoms to clump together. The minute clumps serve as nuclei on which crystals precipitate. These crystals originate from compounds in the glass itself, such as lithium metasilicate, or from additives, such as sodium fluoride. If sufficient numbers of fairly large crystals form, the glass becomes a translucent or opaque opal. Masked regions remain transparent and glassy.

PHOTOCHROMIC GLASS finds primary use as lenses that darken outdoors. Here only one of the lenses has been exposed to ultraviolet light.





ter left). *More commonly, however, the glasses find applications as glass-ceramic channels—here, 0.38 millimeter in diameter with a tolerance of 0.01 millimeter—that align the pins of dot-matrix printers* (center right) *and as cell sheets that help to separate lighted pixels in gas-discharge displays* (far right).

Photosensitive glasses have become more than just a new photographic medium, as they were described when first discovered. In the 1980s Nicholas F. Borrelli and David L. Morse of Corning found that the chemical process can lead to the manufacture of microlenses. Lithium metasilicate opal shrinks and becomes denser than the glass from which it formed. The effect occurs while the glass is heated to its softening point; the glass deforms to accommodate the shrinkage. By carefully manipulating the temperature and ultraviolet exposure, workers are able to control the deformation of the glassy regions and thus regulate the optical power of the lenses.

In 1984 Dennis W. Smith and I found that the different properties of opal and glass enable the direct formation of electrical circuitry on glass. Immersing a sample with an opal pattern in a molten salt bath containing silver ions and heating it in a hydrogen atmosphere forms a conducting silver film on the opal regions. The film that forms on the glassy regions remains nonconducting. This phenonemon has yet to be commercially exploited, but it may find application as, for example, a connection between integrated circuits.

Photosensitive glasses can also form highly detailed, complex patterns that are used not only as holiday ornaments but also as spacers in photomultiplier tubes, cell sheets in gas-discharge displays, and charge plates and nozzles in ink-jet printers. Because lithium metasilicate opal is much more soluble in acid than is the glass from which it forms, placing the sample in a bath of dilute hydrofluoric acid leaches out the opal regions. This process leaves behind a complex glass part of high precision.

Using the additive sodium fluoride will result in photosensitive effects different from that achieved by lithium metasilicate. When workers mix sodium fluoride with sodium bromide, they can generate in glass deep, beautiful colors from the entire spectrum.

The initial heat treatment causes sodium fluoride crystals to form on silver nuclei. The blocky crystals serve as bases from which pyramid-shaped crystallites of sodium bromide grow during continued heat treatment. The resulting crystals tend, however, to remain so sparsely distributed and so small that the glass does not become opal but instead retains its transparency. A second, more severe ultraviolet exposure followed by another heat treatment frees additional quantities of silver. This silver precipitates out onto the tips of the sodium bromide pyramids in long, needlelike specks, which are large enough to absorb visible light. The needles are all similar; thus, they absorb the same fairly narrow band of wavelengths and impart a brilliant color to the glass. Varying the ultraviolet intensity results in different dimensions of the silver needles, so that many colors appear throughout the glass. Such colored photosensitive glasses are found in color filters, dinnerware and art glass.

Impermeable, hard, inert, refractory and insulating, photochromic and photosensitive glasses are uniquely useful. Since their invention, they have proved to be the ideal solution for a host of technological problems.

FURTHER READING

FULL-COLOR PHOTOSENSITIVE GLASS. S. Donald Stookey, George H. Beall and Joseph E. Pierson in *Journal of Applied Physics*, Vol. 49, No. 10, pages 5114–5123; October 1978.

DISCONTINUOUS/CONTINUOUS METAL FILMS GROWN ON PHOTOSENSITIVE GLASS. D. M. Trotter, Jr., and D. W. Smith in *Applied Physics Letters*, Vol. 45, No. 1, pages 112–114; July 1, 1984.

PHOTOCHROMATIC GLASSES. R. J. Araujo and N. F. Borrelli in *Optical Properties of Glass*. Edited by D. R. Uhlmann and N. J. Kreidl. Academic Press, 1990.

How Dinosaurs Ran

Did the giants of the Mesozoic period lumber sluggishly, or were they formidable running machines? Techniques borrowed from modern physics and engineering may give us the answer

by R. McNeill Alexander

Elephants do not prance and leap about like gazelles, and they cannot do so because of their size. Imagine how dinosaurs, some of them much more massive than any elephant, must have moved. Perhaps their legs were too weak to support their enormous weight, so they lived in lakes and depended on water to buoy them up, as some old pictures suggest. Or perhaps their limbs were strong enough to support them and they walked and ran like elephants—or some of the more athletic modern animals—despite their size.

Because the dinosaurs are extinct, we cannot confirm or refute any of these possibilities by direct observation, but fairly convincing answers can be obtained by applying methods taken from physics and engineering. Using the art of shipbuilding, the study of forces and stresses on structures and skeletons and the observation of wildlife, we can understand some of the physical rules that may have governed dinosaur movement.

We can combine this mechanical approach with what is known from the fossil record, both from bones and from dinosaur tracks, to describe what may have been the agility or sluggishness of these giant animals. The picture that emerges can tell us whether *Triceratops*—the horned quadruped—galloped or shuffled and whether *Tyrannosaurus*—the bipedal carnivorous king—could outmaneuver or outrun his *Triceratops* prey.

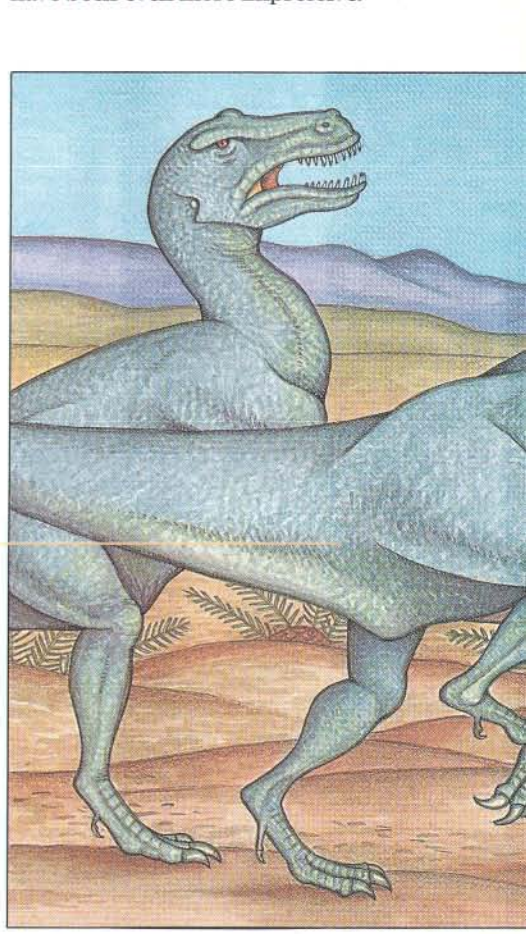
R. MCNEILL ALEXANDER is professor of zoology at the University of Leeds in England. He researches and writes about the mechanics of human and animal movement and is particularly interested in running and jumping, which he studies by film analysis, force recording and mathematical modeling. (Alexander has been known to include his family in some of his experiments—along with jellyfish, kangaroos, frogs and dogs.) He also studies the strengths of animal parts in relation to the forces they have to withstand as well as the consequences of size differences for their structure and movement.

We know dinosaurs principally by their skeletons, from which we can measure their heights and lengths. But this information does not provide the weight that these skeletons, particularly their limbs, had to support. Nor do the measurements of length and height help us make comparisons between dinosaurs of different shapes. Unfortunately, we lack the most generally useful measure of size: body mass.

Most of the weight of living dinosaurs was skin, flesh and guts that decayed long ago, so today their body mass can only be estimated. Such calculations can be done by measuring scale models of the dinosaurs as they are thought to have been in life—suitable plastic models can be bought in many museums.

Using Archimedes' principle (immersing a model in water and measuring the amount of water displaced), we can determine the volume of plastic replicas. This measure can then be scaled up to approximate the volumes of the real creatures. Some models are made to a scale of one fortieth, so their volumes have to be multiplied by $40 \times 40 \times 40$ (length times height times width), or 64,000, to get the volumes of the actual dinosaurs. Multiplying this staggering volume, in turn, by 1,000 kilograms per cubic meter (the density of water and, approximately, of crocodiles and mammals) gives the dinosaurs' likely masses.

These measurements suggest a mass of more than seven metric tons for *Tyrannosaurus*, the largest-known flesh-eating dinosaur. (A metric ton is 2,205 pounds.) This weight is 10 times that of a fully grown male polar bear, the biggest modern land-living predator. *Brachiosaurus*, the largest plant-eating dinosaur of which there exists a reasonably complete skeleton, was even more gargantuan. It seems to have weighed about 50 metric tons, 10 times the mass of a mature male African elephant and about equivalent to that of an average sperm whale. Although *Brachiosaurus* was indeed a magnificent animal, standing 13 meters high, or well over twice as tall as an adult giraffe, it may not have been the biggest dinosaur. *Supersaurus* and *Ultrasaurus*, known only by a few bones each, may have been even more impressive.



***TYRANNOSAURUS*, a carnivore, pursues the three-horned plant-eating *Triceratops*. Was the chase quick and lively, or did these two dinosaurs plod along?**

The problem of support faced by very large land animals was first addressed by Galileo in the early 1600s, when he theorized about the relation of size to strength and structure. Consider two animals of different sizes that are geometrically similar—meaning that if the larger is, for example, twice as long as the smaller animal, it is also twice as wide and twice as high. In that case, the bigger animal has two (length) times two (height) times two (width), or eight, times the volume of the smaller one ($2 \times 2 \times 2 = 8$). And assuming that they are made of the same substances, the larger creature outweighs its smaller counterpart eight times.

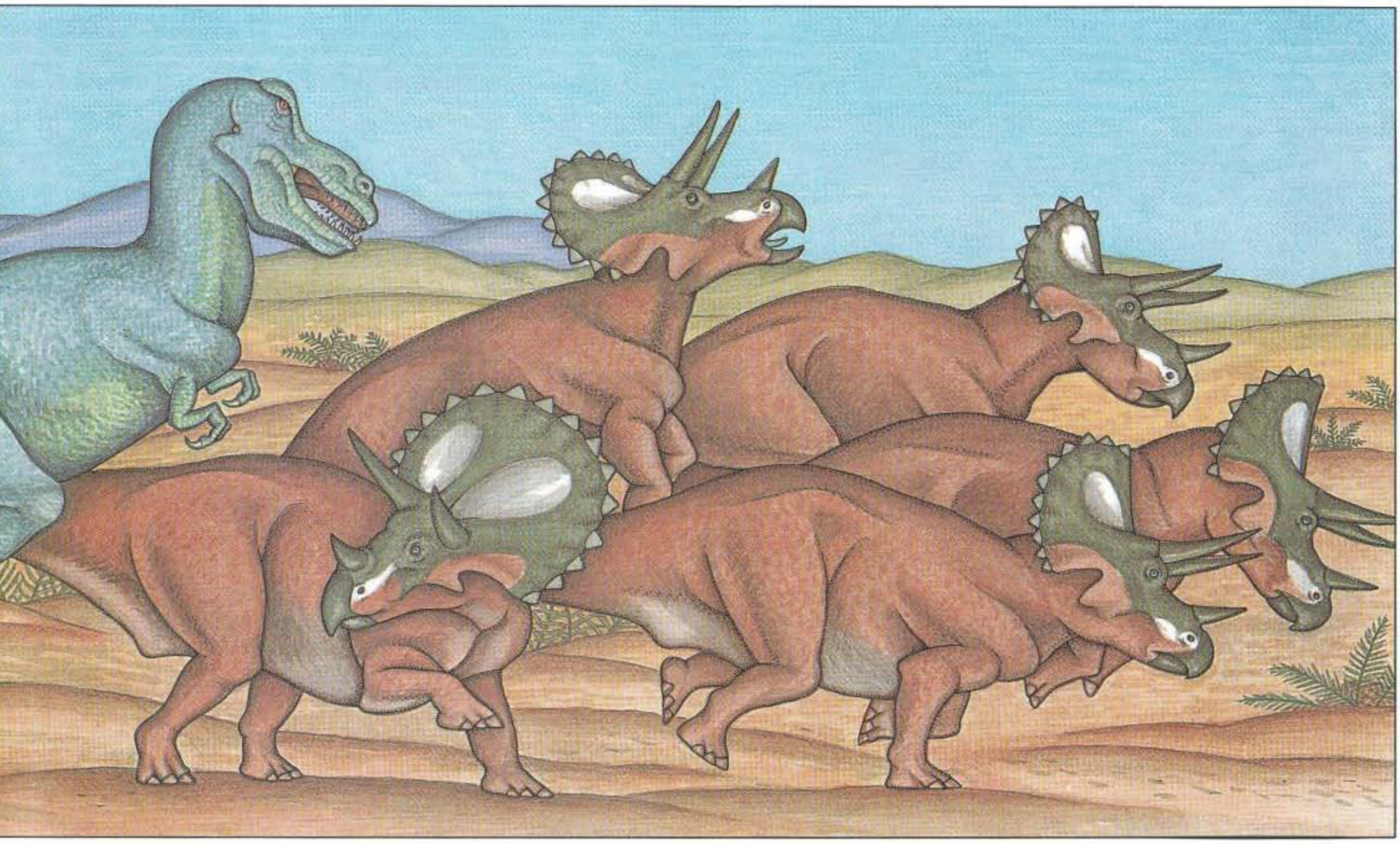
But there is a complication. Although the volume of the larger animal is eight times greater, the strength of its legs increases only by a factor of four. Because leg strength is proportional to the area of the cross section of the limb, one leg would be only two (length) times two (width), or four, times stronger. (An increase in height would have no effect on the cross section.) In other words, eight times the weight would have to be carried by only four times the strength. So, as Galileo noted, if an animal becomes progressively bigger without changing its shape, it must eventually reach a size at which it is incapable of supporting itself.

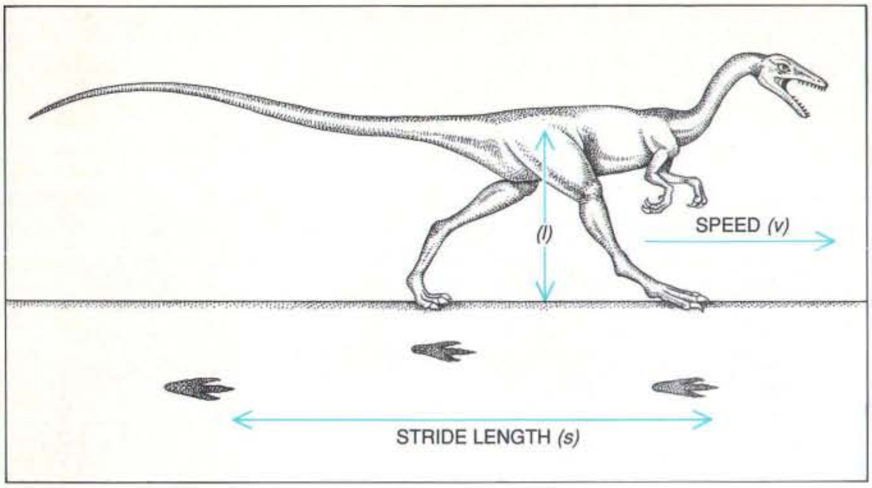
To apply this observation about support to the movement of dinosaurs, we need information about modern animals—particularly about the effects of size differences on their movement. I had been researching the mechanics of running and jumping in frogs, dogs, kangaroos and other animals for several years, when I was invited to speak at a conference about the impact of size on animal locomotion. So I set out to develop a theory of running that would account for size differences and soon noticed that an expression kept cropping up in my equations. This was v^2/gl, where v is running speed, g is the acceleration of gravity (9.8 meters per second squared) and l is leg length. This expression held the key that eventually made it possible to assess the athletic abilities of dinosaurs and even to calculate their running speeds.

Although I did not realize it at first, v^2/gl falls into a class of expressions first used by a 19th-century naval architect named William Froude. Froude wanted to estimate the power needed to propel newly designed ships by making tests on small scale models before the ship itself was built. (In doing so, he hoped to avoid expensive mistakes.) Since the bow wave that is pushed in front of a ship causes much of the resistance to the vessel, Froude realized that the height of the bow wave in his models must be proportionate to that of a real wave against the prow of a real ship. To achieve this wave size, he showed that the model must be tested at a specific speed, giving rise to an expression now named after him: a Froude number.

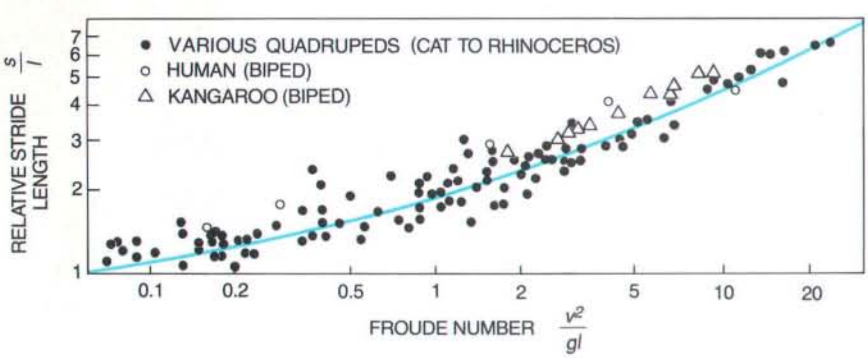
In the realm of shipbuilding, the model's Froude number (v^2/gl) must be the same as the Froude number for the real ship traveling at its normal speed in order for the test to be useful. In Froude's calculations, of course, l was not leg length but the length of the hull of the ship.

Froude's experiment illustrated a special case for a more general rule—if it had not, it would have been of no use to me, since ships are not much like dinosaurs. The rule involves the concept of dynamic similarity, which can be thought of as an extension of geometric similarity. As we saw, two shapes are geometrically similar if one can become identical to the other by a uniform change in the scale of length. (Height and width, of course, have to be scaled up in proportion.) Likewise, two motions are dynamically similar if one can become identical to the other by uniform changes in the scales of





STRIDE LENGTH is the distance between two successive prints from the same foot. *Compsognathus*, a carnivore the size of a contemporary chicken, is depicted here.



FROUDE NUMBERS for kangaroos, humans and quadrupeds, such as rhinoceroses, are plotted against the animals' relative stride lengths. The numbers increase logarithmically—so that the difference between Froude numbers 0.1 and 20 is clear.

length, time and force. Two animals of different sizes, for example, may be dynamically similar if they both are trotting or galloping.

The general rule that applies to bow waves, running animals and all other systems involving gravity reveals that dynamic similarity is possible only when Froude numbers are equal. Looking at the energy of these systems clarifies this point. In the rise and fall of a wave, or of a limb, energy is converted back and forth between kinetic and potential forms. Two systems can be dynamically similar only if they have the same ratio of kinetic energy (equal to $1/2\ mv^2$ for a body of mass m moving at speed v) to potential energy (equal to mgh, where h is height). This ratio is $v^2/2gh$, proportional to a Froude number.

With the formula in hand, I was able to extend the observations from shipbuilding to other aspects of dynamic similarity. I hypothesized that geometrically similar animals of different sizes would run in approximately dynamically similar fashion whenever their speeds made their Froude numbers equal. Precise dynamic similarity could not be expected, because animals of different sizes are not quite the same shape, and strict dynamic similarity requires strict geometric similarity.

The theory proved generally true. For instance, one of the predictions of the hypothesis was that animals of different sizes would use similar gaits when their Froude numbers are equal. Indeed, ferrets and rhinoceroses change from trotting to galloping at very different speeds, but in each case the Froude number is about the same. Ferrets alter their gait when they start to travel at 1.5 meters per second. Because their hip height is 0.09 meters, the expression of the Froude number is 1.5 squared divided by 0.09 times g (or 9.8 meters per second squared). The result is 2.55. Similarly, a rhinoceros changes from a trot to a gallop at 5.5 meters per second, and its hip height is 1.2 meters—when the calculations are made, the Froude number is 2.57.

Another prediction involved stride length, the distance between successive footprints of the same foot. The faster animals run, the longer their stride. The hypothesis held that when their Froude numbers were equal, different animals would take strides in the same proportion to their leg length—because of their dynamic similarity. Therefore, a graph of relative stride length (that is, stride length divided by leg length) plotted against Froude number should be the same for cats and camels, ferrets and rhinoceroses [*see lower illustration at left*].

This relationship, however, turns out to be true only for mammals the size of a house cat and larger. It does not hold as well for smaller mammals such as rats because they run in an often peculiar, crouched manner, quite different from the bigger mammals. The graph also shows that the relation between relative stride length and Froude number is similar for bipeds such as people and kangaroos and for quadrupeds such as dogs and horses—demonstrating that these two forms of movement share some dynamic similarity. (To that extent, you run the way the hindquarters of a small pony move.)

The graph proved to be even more illuminating because it soon occurred to me that I could use it to estimate the speed of dinosaurs from the stride lengths shown by their footprints. Rather surprisingly, great numbers of dinosaur footprints have survived as impressions in mud that turned to stone [see "The Footprints of Extinct Animals," by David J. Mossman and William A. S. Sarjeant; SCIENTIFIC AMERICAN, March 1983]. These tracks show that dinosaurs walked with their feet directly under their bodies like mammals and birds, not sprawled out to either side in the manner of modern reptiles. Consequently, the relation between relative stride length and Froude number for mammals should also apply to dinosaurs. In contrast, estimating dinosaur speed by comparing these giants with modern reptiles would not be appropriate.

The largest-known footprints, with hind feet measuring 1.3 meters in diameter, have been found in Spain. Tracks of slightly smaller prints lie in other parts of the world—even in Yorkshire, England. The best known of the smaller prints—measuring 0.9 to 1.0 meter in diameter—were discovered in Texas. The size and shape of these footprints suggest that sauropods, huge long-necked, long-tailed herbivores, roamed this region. Three-toed bipeds similar to *Tyrannosaurus*

made other tracks nearby. One famous trail shows that these two Texans met: a sauropod and a tyrannosaurlike biped traveled along the same path [*see illustration on this page*]. Was this dramatic chase leading to a kill? Were the animals heaving themselves along slowly and with difficulty, or did they rush past at earth-shaking speeds?

Using Froude numbers and stride lengths measured from these and similar trackways, I made my first efforts to infer the speed of dinosaurs. Since leg length can be estimated from the size of the footprints—footprints should be about one quarter of leg length—relative stride length could also be calculated. And once relative stride length was established, I used the graph to find the corresponding Froude numbers. Then, armed with the leg length, I could calculate speed. The results may not be very accurate, however, because the data points on the graph show a fair amount of scatter and because we are using data from modern animals to estimate speeds for dinosaurs.

By this account, the speed of the large dinosaurs was unimpressive. All known footprints of large sauropods seem to show speeds of about one meter per second, a slow walking speed for humans that seems painfully slow for animals with three-meter hind legs. None of the footprints of very large bipedal dinosaurs show speeds above 2.2 meters per second, the pace of a fast human walk.

Although most of the footprints of big dinosaurs seem to show walking speeds, many footprints of smaller ones record running. The fastest tracks were made in Texas by a biped of probably a little more than half a metric ton, roughly the mass of a racehorse, and by another somewhat smaller one. Both sets of footprints indicate a speed of 12 meters per second, which is higher than the peak speed of 11 meters per second reached by the best human sprinters—but well below the speeds of 15 to 17 meters per second at which horse races are generally won.

The lack of running footprints of huge dinosaurs, however, does not show that they could not run but merely that they usually walked—at least on surfaces in which footprints were likely to be preserved. If you were to go out on a snowy day, for example, and measure human footprints, you would probably find only short strides, indicating walking speeds, but you would be wrong to conclude that people cannot run. Clearly, a different approach was needed to judge how athletically large dinosaurs could have moved when they really tried.

The faster an animal runs, the greater the forces its feet exert on the ground and the stronger its legs need to be. The reason is that at higher speeds each foot is on the ground for a smaller fraction of the stride, so it has to exert peak force—the maximum force that occurs while the foot is on the ground—to make a complete stride balance and carry the body weight. For instance, the peak force on a human foot rises from simple body weight during slow walking to 3.5 times the body weight in sprinting. Fast running and athletic behavior require strong bones.

Examining the forces exerted by the feet of various creatures led to a greater understanding of dinosaur motion. My colleagues and I used a force plate (a pressure-sensitive panel embedded in the floor) to measure the forces on the feet of people, dogs, a sheep and a kangaroo as they walked, ran, hopped or jumped. These results, together with films of animals running and anatomical data, enabled us to calculate the stresses that these different movements imposed on the animals' leg bones.

A few assumptions also allowed us to determine the stresses in the bones of wild animals such as buffalo that we could film but not bring into the laboratory. We could have applied the same

DINOSAUR TRACKS provide a record of stride length and speed. A small, three-toed carnivore may have pursued a larger sauropod along this Texan trail. This pair of footprints was discovered by Ronald T. Bird at Paluxy Creek in 1944.

approach to dinosaurs to find out how well their bones would have withstood the stresses of running, but reconstructing their patterns of movement would have involved elaborate calculations and a good deal of imagination. I preferred a quicker and easier way, again using the concept of dynamic similarity and insights from structural engineering.

Forces act on the ends of bones (at the joints) setting up stresses in the bone shaft. These forces can be broken down into their components: axial force (F_{ax}) acts along the long axis of the bone, and transverse force (F_{trans}) acts at right angles to it [*see box on opposite page*]. Taken alone, F_{ax} sets up a uniform stress $-F_{ax}/A$, where A is a cross section of the bone and the minus sign indicates a compressive stress. Added to this force are the stresses caused by F_{trans}. These transverse stresses vary across the thickness of the bone from $-F_{trans}x/Z$ at one end of the bone to $+F_{trans}x/Z$ at the other. In these expressions, x is the distance of the cross section from the end of the bone, and Z is the section modulus, a geometric property of the cross section. (Engineering textbooks explain how to determine the section modulus, for those interested.)

Too great a stress will break a bone. Calculations for the leg bones of running and jumping modern animals showed that the stress $F_{trans}x/Z$ was generally much greater than the stress F_{ax}/A. This difference tells us that transverse forces are a much more serious threat to bones than axial ones: it is much easier to break a stick or a bone or any other long, thin bar by transverse forces than it is to break it by axial ones. Therefore, since we need only rough estimates of bone stresses, we can ignore the less important F_{ax} and consider only F_{trans}.

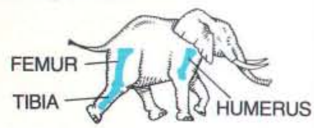
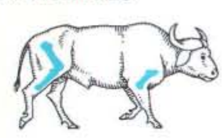
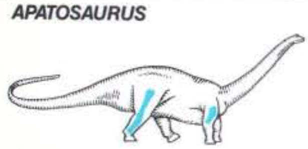
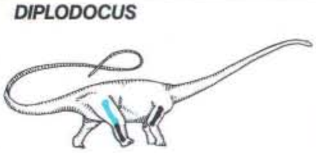
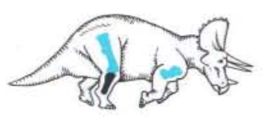
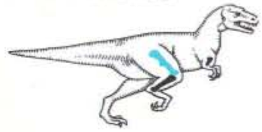
For animals running in dynamically similar fashion, all the forces acting on the bones are proportional to body weight, W, because the bones have to support that weight. This rule includes transverse forces on the leg bones: the stresses they cause ($F_{trans}x/Z$) are proportional to Wx/Z. Now imagine two similar animals of different sizes moving in a dynamically similar fashion. The stresses in the leg bones will be less in whichever animal has lower values of Wx/Z, indicating that its bones are strong enough to do more athletic things than the other animal's bones.

By changing the expression just slightly, we can arrive at a value for a strength indicator that is more straightforward to use: the reciprocal Z/Wx. Thus, the greater the value of Z/Wx is for the bones of the animal, the more athletic we can expect that animal to be. To be even more precise, we can use Z/aWx, where a is the fraction of body weight supported by the forelegs or hind legs, as appropriate. Including this weight distribution enables us to make meaningful comparisons between, for example, elephants and the brontosaur *Apatosaurus*—a massive herbivore dinosaur—since elephants carry the major part of their weight on the forelegs and *Apatosaurus* on the hind legs.

Once we had determined a strength indicator for the principal leg bones of large modern mammals and of dinosaurs, we were able to conclude that a dinosaur could have been as athletic as any modern mammal with similar strength indicators. This conclusion depended on the assumption that the bones of different animals can stand about the same stresses. And tests on samples of bones from birds and mammals show that this is indeed roughly true. We cannot check these stresses on dinosaur bones, however, since their properties have been altered by decay and fossilization.

The conclusion that high strength indicators imply agility also depended on the assumption that evolution has adjusted the strengths of bones of different animals to give them equal safety factors. The safety factor of a structure is described as the force needed to break it, divided by the maximum force expected to act on it in normal use. (Engineers generally design structures with safety factors of two or more to make disaster unlikely.)

Bone Strength in Large Animals

	BODY MASS (METRIC TONS)	STRENGTH INDICATOR FEMUR	TIBIA	HUMERUS
AFRICAN ELEPHANT 	2.5	7	9	11
AFRICAN BUFFALO 	0.5	22	27	21
APATOSAURUS 	34	9	6	14
DIPLODOCUS 	12–19	3–5	NO DATA	NO DATA
TRICERATOPS 	6–9	13–19	NO DATA	14–22
TYRANNOSAURUS 	7.5	9	NO DATA	NO DATA

The process of arriving at the values for the strength indicators for the dinosaurs is not as straightforward as the conclusions. The first step is to calculate the value of the section modulus, or Z, of their bones. To do this, I needed detailed measurements of cross sections at known distances, x, from the end of the bones. (The section modulus, although complicated to arrive at, simply takes account of the areas of all parts of the cross section and their distances from the middle of the bone, or girder.) I was able to obtain values for x from accurate drawings published by earlier paleontologists.

Now all I was missing to make my calculation of the strength indicators complete was the weight, W, of the dinosaur and, if it was a quadruped, the fraction, a, of that weight carried by each pair of feet. (I have already explained how dinosaur weights are estimated from the volumes of models.)

To calculate the distribution of weight between the feet, I had to find the center of gravity of each dinosaur. By suspending each model from a thread—first from its nose so that it hung vertically and then from its back so it hung horizontally—I could arrive at this center. Each model was photographed in both positions, and care was taken to ensure that the axis of the camera was perpendicular to the plane from which the scaled plastic replica hung. (For such an exercise, monofilament nylon is superior to spun thread, which unwinds, thus making the model spin.) Because the dinosaurs hung with their centers of gravity directly below the point of support, the thread pointed toward the center of gravity in each photograph. By superimposing two photographs, I could discover the center of gravity.

One slight complication in this technique was easily remedied. Unlike real dinosaurs, which would have had air in their lungs and dense bones running through less dense flesh, the models were solid plastic of uniform density. In the real animals the bones were distributed along the whole length of the body, so they probably did not affect the position of the center of gravity too much—but the air was all near the front end of the trunk. I compensated for these different densities either by calculation (assuming lungs of about the same fraction of body volume as in modern reptiles and mammals) or by boring a hole representing the approximate volume where the lungs would have been.

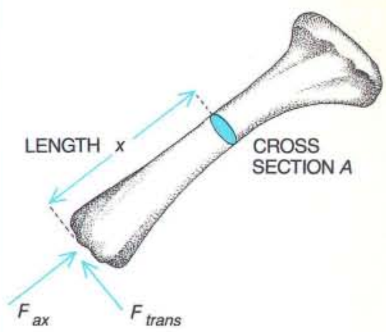
Correction for the lungs moved the center of mass only a little bit, less than 20 centimeters in a 20-meter sauropod. Once I knew, or thought I knew, the dinosaur's center of mass, I could divide its weight appropriately between the forefeet and hind feet. (For example, if the center of mass was two meters behind the front feet and one meter in front of the hind ones, the hind feet would have carried two thirds of the weight.)

All these procedures and data, taken together, supplied the information we needed to calculate the strength indicators Z/aWx for different dinosaur limbs [*see table on opposite page*]. Remember that obtaining larger values from this formula implies more agility for the animal. The computations enabled us to compare elephants with *Apatosaurus*, the very large sauropod dinosaur commonly known as the brontosaur. The strength indicators ranging from seven to 11 for the various leg bones of an African elephant are not very different from the values of six to 14 for *Apatosaurus*—whose bones were bigger but similarly proportioned.

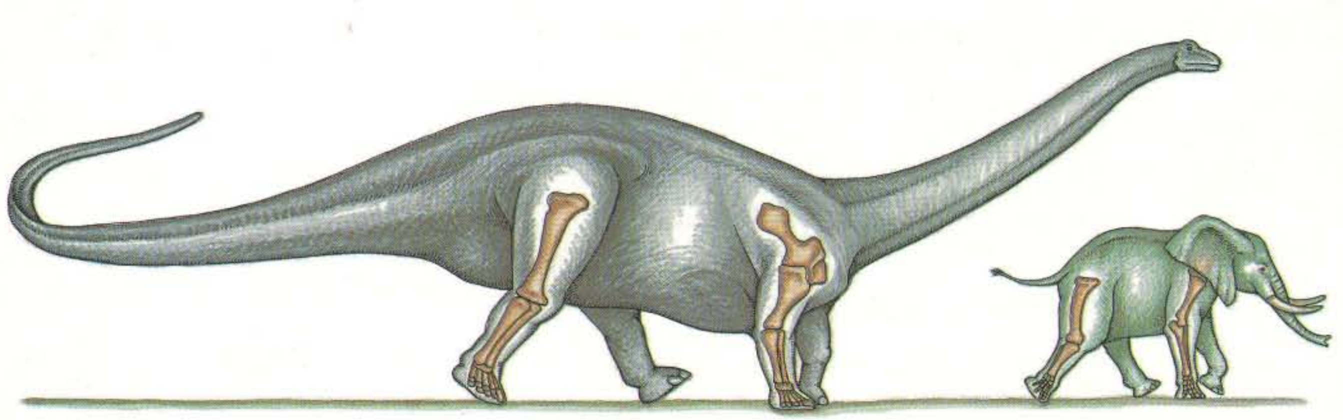
This comparison suggests that despite its huge size, *Apatosaurus* could have been about as athletic as an elephant. Elephants can manage a slow run but cannot gallop or jump (a narrow ditch is enough to keep them in a zoo enclosure). It seems likely that the same was true of *Apatosaurus*. Unfortunately, it is difficult to be specific about speeds because most published elephant speeds are subjective estimates, and anyone who has had an alarming experience with an elephant is likely to exaggerate. One film of a young elephant running shows a speed

How Forces Act on Bones

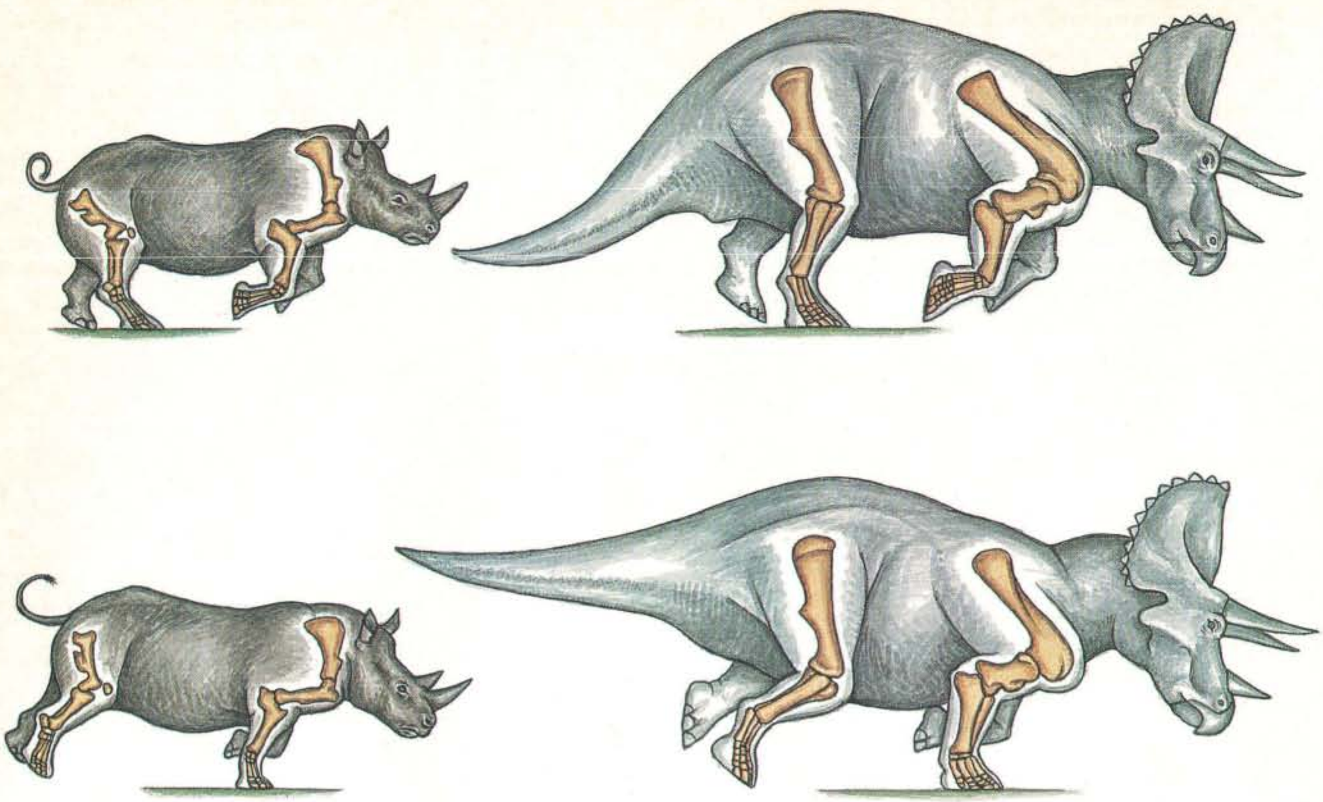
When an animal moves, forces are exerted on the ends of its bones and cause stresses in the bone shaft. These forces can be broken down into different components: the axial force (F_{ax}) and the transverse force (F_{trans}). In the



cross section (*area A*) of the diagram, F_{ax} produces a stress, $-F_{ax}/A$ (the stress is negative because it is compressive). This stress is weaker than that stress produced by F_{trans}: $F_{trans}x/Z$. (Z represents the section modulus, an engineering term, which describes properties of the cross section.)



ELEPHANT AND *APATOSAURUS*, a plant-eating dinosaur, both have approximately the same strength indicators for their leg bones. This similarity means that the extinct giant probably ran much the same way our contemporary does.



TRICERATOPS may have moved like the White rhinoceros, a modern, horned herbivore. The White rhinoceros here, which was sketched from a film, is galloping at seven meters per second—which is about the speed of a fast human run.

of about five meters per second (a slowish speed for a human runner). To reach the same Froude number on its longer legs, *Apatosaurus* would have had to run at seven meters per second—slightly brisker than an elephant.

Diplodocus, a more slender sauropod, seems to have a lower strength indicator, at least for the femur (the only bone for which I found a cross section). Therefore, it was probably less agile, able to walk on land with no need to be buoyed up by water but possibly not able to run. Unfortunately, considerable doubt exists about the mass both of *Diplodocus* and of the well-known horned *Triceratops* because measurements have been made in each case on a skinny model and a stout one—it seems uncertain which model is more accurate.

The strength indicators for the *Triceratops* are higher than for either of the two sauropods and fall between elephant values and values for more athletic animals such as the African buffalo. These findings suggest that *Triceratops* may have been more athletic than elephants, possibly able to gallop like buffalo and rhinoceroses. I have a film of an adult White rhinoceros of about two tons, galloping in a large zoo enclosure at seven meters per second while being pursued by a vehicle. For *Triceratops* to have reached that same Froude number—although it is doubtful whether it could have—it would have had to travel at nine meters per second.

The conclusions reached for *Apatosaurus, Diplodocus* and *Triceratops* are already tentative, and we must be even more cautious about *Tyrannosaurus* because all modern bipeds are so very much smaller. Indeed, no contemporary biped moves its legs in a way that seems likely for its structure. We can only note that the strength indicator of a *Tyrannosaurus* femur is low, in the elephant range.

The calculations that allowed me to assess the agility of dinosaurs are firmly based in physics and engineering. Especially useful to me was the concept of dynamic similarity, which had its origins in shipbuilding but has become immensely important in aerodynamics, heat engineering and other branches of physical science. And the theory of stresses in beams (the section modulus), which is constantly used by structural engineers, provided further insights.

Both sets of theory enabled me to arrive at dinosaur strengths and speeds, but it would be foolish to claim that the calculations are accurate. I hope I have made clear some of the sources of error. These estimates give the impression that although large dinosaurs walked slowly, most were capable of quite a quick run and none needed to live in water or to rely on buoyancy for support.

The evidence from footprints suggests that if we had been alive at the time (and had had the nerve), we could have strolled alongside a walking sauropod or tyrannosaur, keeping up with it without difficulty. The calculations derived from bone dimensions suggest that large sauropods may have been as agile as elephants and that *Triceratops* may have been a little more athletic. I think I am probably fast enough to outrun a pursuing tyrannosaur, but, perhaps fortunately, I am unlikely to have to try.

FURTHER READING

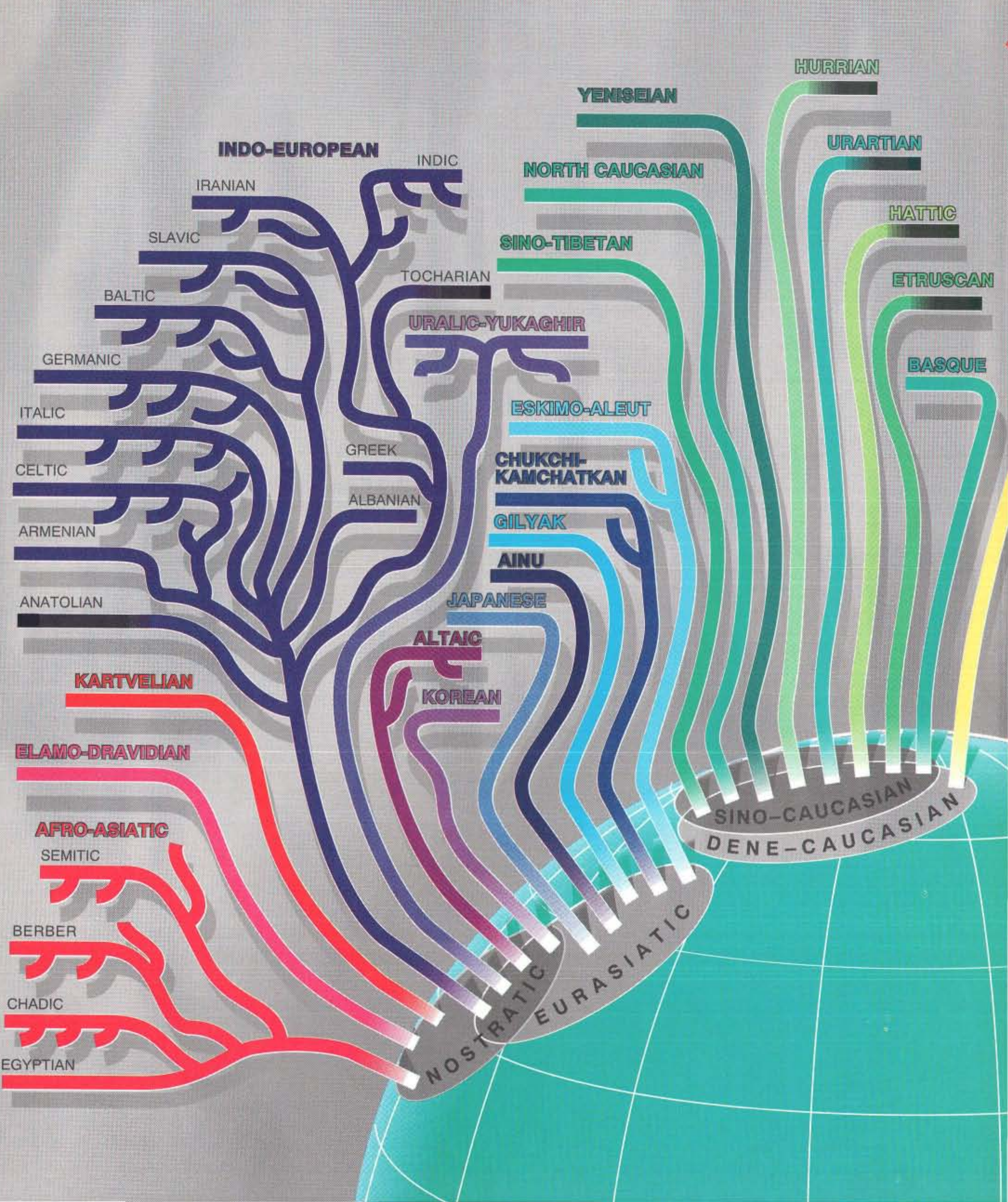
MECHANICS OF POSTURE AND GAIT OF SOME LARGE DINOSAURS. R. McN. Alexander in *Zoological Journal of the Linnean Society,* Vol. 83, No. 1, pages 1–25; February 1985.

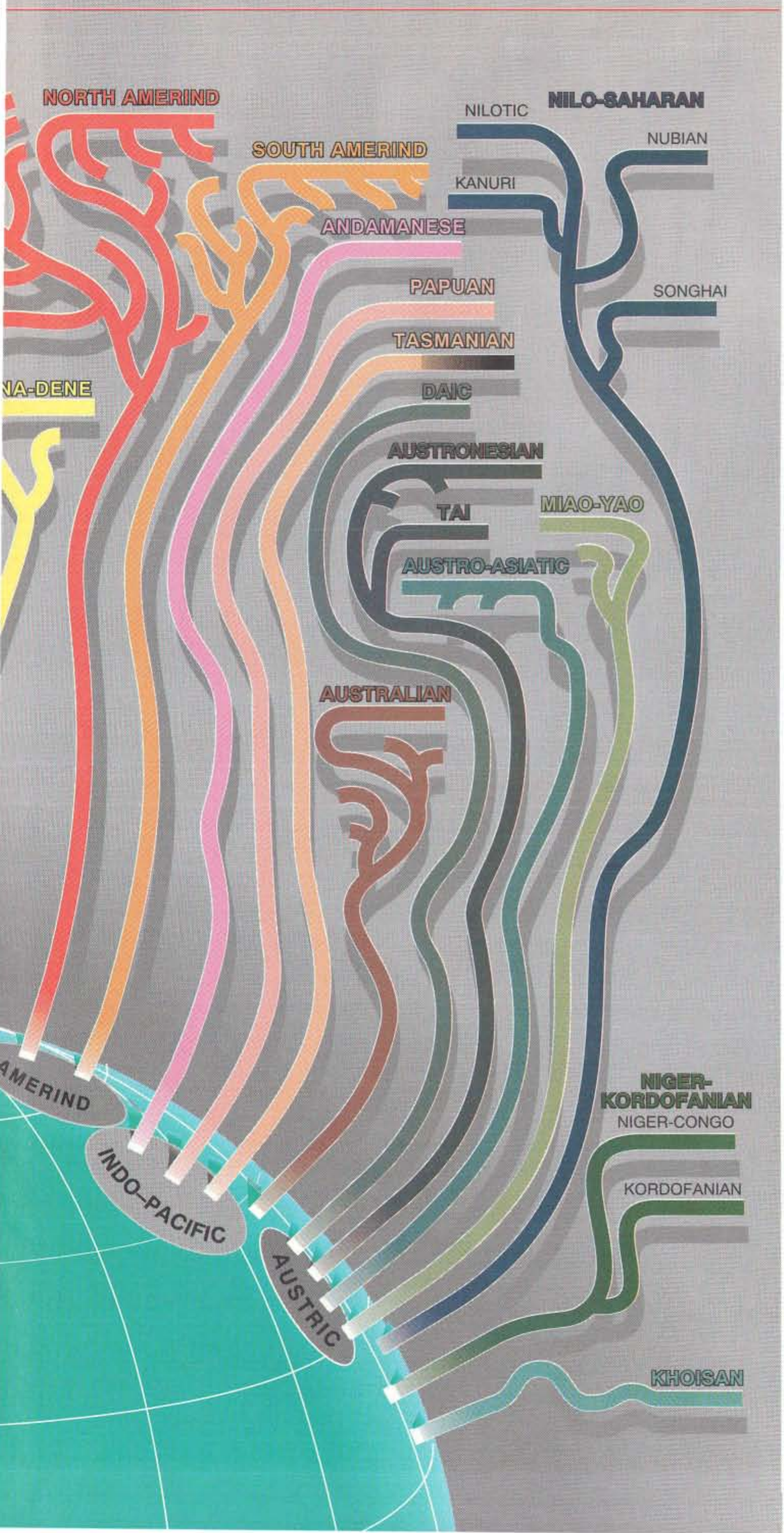
DYNAMICS OF DINOSAURS AND OTHER EXTINCT GIANTS. R. McNeill Alexander. Columbia University Press, 1989.

DINOSAUR TRACKS. Tony Thulborn. Chapman and Hall, 1990.

HARD WORDS

by Philip E. Ross, *staff writer*





How deeply can language be traced? Radical linguists look back to the Stone Age. Traditionalists disagree.

In the beginning there was the word. Ask Merritt Ruhlen what it was and he will reply: *tik*. That simple monosyllable was what prehistoric man may have named finger. Ruhlen says it comes down to English as *toe* and to Latin as *digit*. He refuses to put a date on the root, although others argue, on genetic evidence, that it goes back perhaps 100,000 years.

Ruhlen, an unaffiliated linguist, is one of a small group of radical researchers who believe they can hear echoes of ancient voices. He thinks it is possible to trace all the world's languages back to a single source spoken in the distant past, a tongue not heard since well before the first horse was broken and the first dog was brought unflinching to the campfire. Others are content to trace language back to roots current perhaps 12,000 to 15,000 years ago, before the development of agriculture.

That all the several thousand languages spoken by the world's four billion people sprang from a common root is a powerful idea indeed. It is the linguistic parallel of the "Eve hypothesis" formulated by Allan C. Wilson, Mark Stoneking and Rebecca L. Cann of the University of California at

GENEALOGICAL CHART relates entire families of languages to proposed superfamilies. The Indo-European family, for example, appears in purple; it includes English in its Germanic branch (top left). *Some Soviet linguists group the family with six others in the controversial Nostratic superfamily* (bottom left). *An alternative superfamily, Eurasiatic, has also been proposed. None of the superfamilies and not all of the familes are accepted as proved by the majority of linguists. Families, such as Etruscan, and subfamilies, such as Anatolian, that have no living languages appear in black.*

Berkeley, who compared samples of DNA to trace the entire human population back to a single woman in Africa perhaps 150,000 years ago. It may even rival the importance of the grand unified theory now being sought by physicists. But, like its counterparts in other disciplines, the monogenesis of language is hard to prove.

For nearly two centuries, scholars have been grouping languages into families, some 200 in all. Some families, known as isolates, contain only a single member: Basque, a language still spoken in the Spanish and French Pyrenees, is the best-known example. But most contain a number of languages whose similarities point to their descent from a common ancestor, the protolanguage of the family. The most ancient protolanguages that all linguists accept are believed to have been spoken about 7,000 years ago.

To comparative linguists who have devoted entire careers to a painstaking study of one group of languages, the idea that the families themselves might be compared and their far more ancient roots discerned seems impossibly ambitious. If the tree of language could in fact be traced to far deeper roots, then the trunks on which most specialists have labored would seem mere twigs on a vaster tree.

Shouted Down

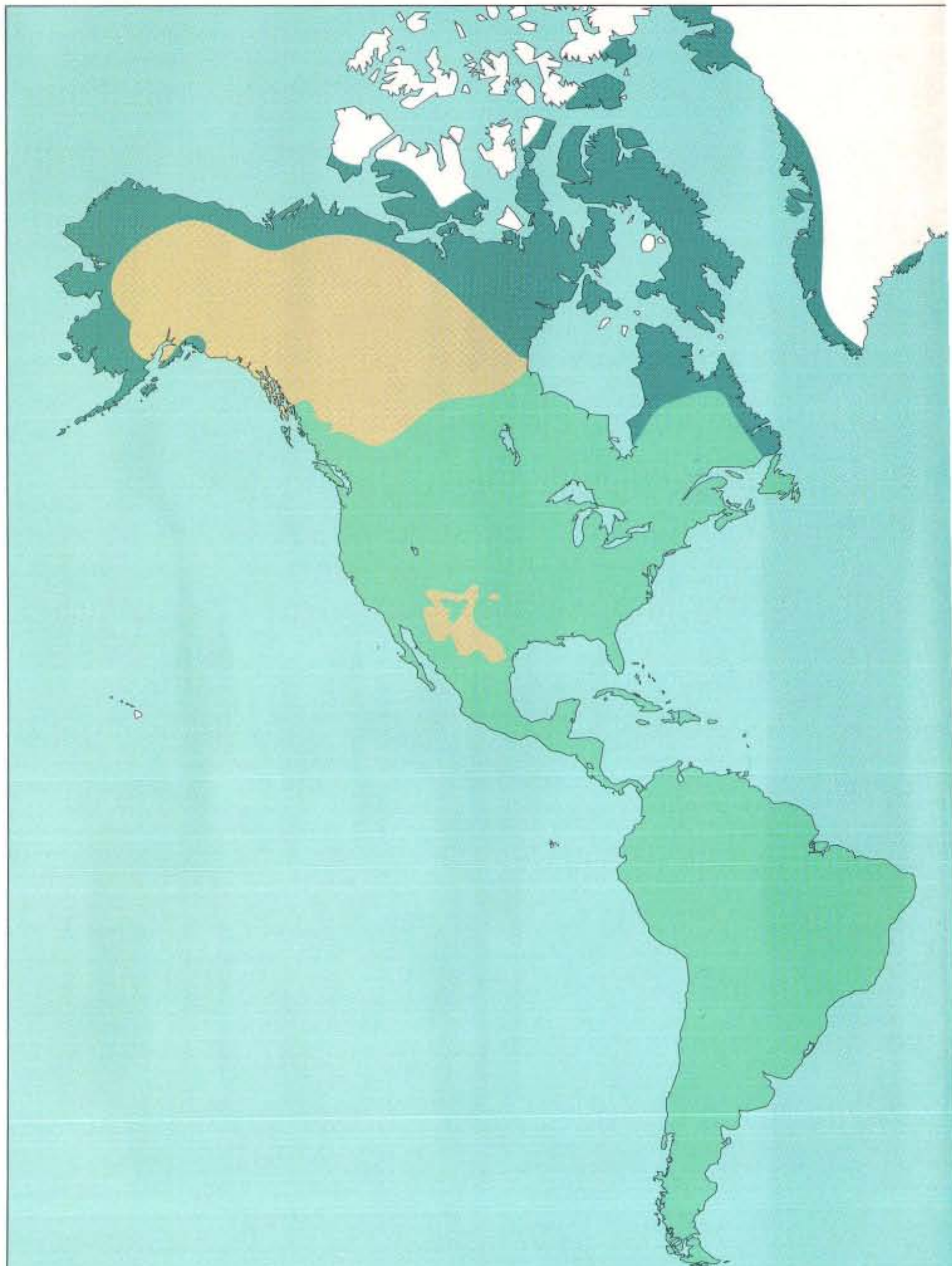
Yet that is precisely what two groups of researchers are doing. The first serious effort to trace the branches of human language back to a 12,000-year-old Neolithic trunk was made nearly 30 years ago by the Soviet linguists Vladislav M. Illych-Svitych and Aaron B. Dolgopolsky. The two scholars, at first working independently, eventually linked six families of languages to trace a hypothetical ancestor that they called Nostratic, derived from the Latin for our (language). Together these families transmit the cultural heritage of three quarters of humankind.

In the U.S., Joseph H. Greenberg of Stanford University began comparing families of languages in the 1950s. His most sweeping work, *Language in the Americas,* was published in 1987. In that book, Greenberg classified the myriad languages of the Americas into three groups, each with its own ancient antecedents. He thus united three quarters of the language families that are known to exist or to have ever existed.

Even before it was published, Greenberg's analysis came under attack. In 1986 a preview of his book appeared in *Current Anthropology.* That inspired Lyle Campbell, who teaches Native American languages at the University of Louisiana, to write that Greenberg's classification should be "shouted down" so as not to mislead anthropologists into wasting time trying to confirm it.

But the strongest opposition to the Nostraticists and to Greenberg and his associates comes from the traditional comparative linguists who specialize in the Indo-European family of languages. This family is thought to be descended from a single speech last spoken in the fourth or fifth millennium B.C.

All specialists agree that by the dawn



Living Language Families of the World

The map gives current ranges of major language families, most of which are generally accepted by linguists. Some, however, are disputed, notably the Amerind family. Wherever recently adopted Indo-European tongues such as English, French or Portuguese would have obscured a whole family, the indigenous family is shown instead.

Western Hemisphere: From Joseph H. Greenberg, Language in the Americas.

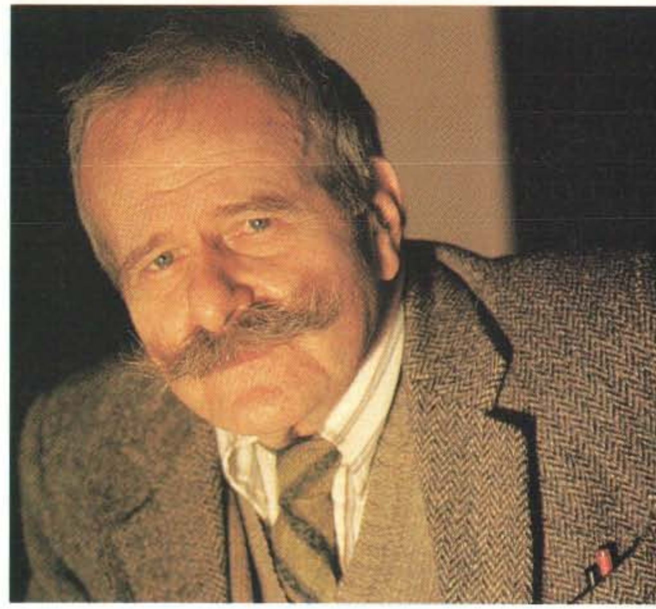
of history, about 4,000 years ago, proto-Indo-European had fragmented into a dozen branches, two of which, Anatolian and Tocharian, left no survivors. By the Middle Ages, Germanic, Italic, Celtic, Baltic, Slavic, Albanian, Greek, Armenian, Iranian and Indic had differentiated into a plethora of modern tongues, several of which have since ridden through the world on the shoulders of conquerors, farmers, merchants and missionaries, becoming the native speech of every second human being.

Indo-European study has its origin in the systematic, if rather fanciful, classifications of the 16th and 17th centuries, when theologians first attempted to prove the biblical account of the confusion of tongues at Babel by tracing all languages to biblical Hebrew. Secular etymologies were tried as early as

Eastern Hemisphere: Adapted from David Crystal, Cambridge Encyclopedia of Language.

Dueling Linguists

Language researchers vehemently dispute how far back one can trace the genealogy of language. Eric P. Hamp of the University of Chicago (*left*) reflects the conservative approach of researchers who reconstruct the history of the Indo-European languages, the best-studied family. He will accept no etymology that does not account for every alteration in sound and meaning that a word has undergone throughout its history. Aaron B. Dolgopolsky of the University of Haifa (*center*) cofounded the study of the proposed Nostratic superfamily 30 years ago, while he was still working in the Soviet Union. Dolgopolsky insists that the 1,600 Nostratic roots he and his colleagues have reconstructed meet Hamp's high standards, even though the etymologies begin where the Indo-Europeanists left off. Joseph H. Greenberg of Stanford University (*right*) represents the most radical school of thought. He groups whole families of languages—most recently the 150-odd families of the New World—into a few superfamilies without bothering to reconstruct the various stages in their development. He says his method of comparing many languages all at once allows him to discern patterns of kinship that more narrowly specialized linguists miss.



the 17th century, notably by Gottfried Wilhelm Leibniz, the co-discoverer of calculus.

But linguists like to date their field to its first great achievement, the argument propounded in 1786 by Sir William Jones, a scholar and jurist. He observed of Sanskrit, Greek and Latin that "no philologer could examine them all three without believing them to have sprung from some common source, which, perhaps, no longer exists."

Jones's immediate successors were the Dane Rasmus Rask and the Germans Franz Bopp and Jacob Grimm, one of the Brothers Grimm of fairy-tale fame. Grimm was the first to call general attention to systematic sound differences in words having similar meanings from different Indo-European groups. He noted that such modern members of the Germanic family as English and German have an *f* and *v* in places where the other groups have a *p*. For example, English and German say *father* and *vater*, whereas Latin has *pater* and Sanskrit gives *pitar-*. A web of such sound shifts, as they are called, was subsequently found to have left marks in many other languages.

The comparative technique attained nearly modern form in the mid-19th century in the work of August Schleicher, who first charted language families as branches on a tree. Schleicher was also the first scholar to attempt to reconstruct languages by inferring their words from later forms, a laborious process that has been likened to triangulation. He even imagined how these words might have produced sentences in combination with grammatical markers, which he also reconstructed.

Historical linguists look for analogies that they then test against a range of criteria to determine whether the analogy is a form that descended directly from the two languages' last common ancestor. Such forms are called cognates. But in language, where change is the rule and words are traded, modified and discarded, reconstructing true meanings is a complex process.

At the foundation of this method is the idea that two languages can be genetically related even if they have no cognates in common. If language A shares cognates with language B, which in turn shares cognates with language C, all three must be kin. Yet one can imagine that the lexicons of A and C might not overlap. Kinship refers to the history of a language, not to its content.

Traditional linguists compare only a few languages at a time, seeking commonalities and reconstructing roots. So far the Indo-Europeanists have reconstructed a vast lexicon of words spoken well before writing was invented. Such reconstructions often tell more about early cultures than can be learned from their material artifacts.

Prehistoric Patriarchs

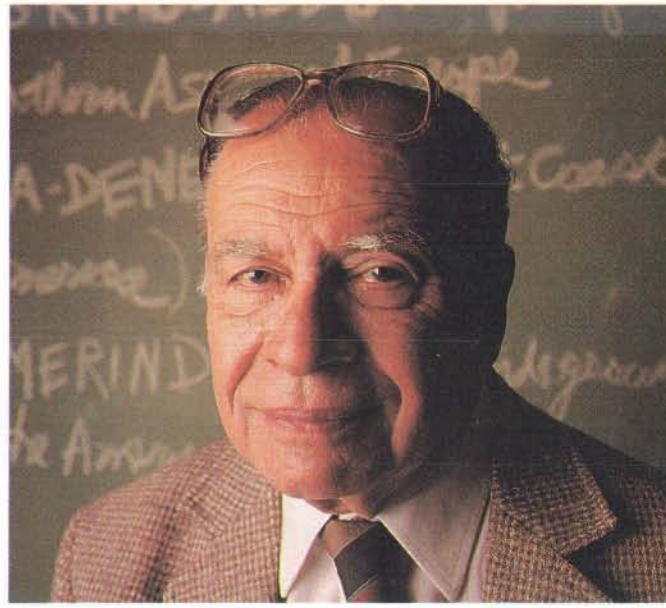
Consider, for example, the reconstructed Indo-European word for father: *p'tēr-*. Linguists modeled this root from derived forms, which comparative studies have shown to signify the male head of a household, as in the Latin *pater familias*. Thus do we know that the speakers of Indo-European formed patriarchal communities.

Similarly, linguists think the Indo-European word for god was *deiw-os*, which later shows up in Latin as *deus*. In combination with *p'tēr-*, it designated the patriarchal god of Indo-European religion: *Dyeu p'ter-*. It survived as *Jupiter* in Latin, *Zeus patēr* in Greek and *Dyaus pitar* in Sanskrit.

By such means can some myths of classical cultures be reconstructed to archetypal patterns, themselves anthropomorphic reflections of the way Indo-Europeans lived. "The reconstructed words *deiw-os* and *Dyeu p'ter* alone tell us more about the conceptual world of the Indo-Europeans than a roomful of graven idols," wrote Calvert Watkins of Harvard University in his introduction to a dictionary of Indo-European roots.

Indeed, the reconstructed proto-Indo-European vocabulary speaks volumes about the lives of those who spoke it. Soviet linguists Thomas V. Gamkrelidze and V. V. Ivanov point out, for example, that the numerous words for domesticated animals, such as dogs, cows and sheep, as well as for crops, such as barley and wheat, indicate that the cultures were primarily agricultural [see "The Early History of Indo-European Languages," by Thomas V. Gamkrelidze and V. V. Ivanov; SCIENTIFIC AMERICAN, March 1990].

Gamkrelidze and Ivanov have also reinterpreted the linguistic evidence to pinpoint the location at which Indo-European arose. Its homeland has been variously placed in the steppes of Russia and the forests of northern Europe. But the two Soviets argue that a number of words seemed to have been borrowed from the non-Indo-European lan-



guages of ancient Mesopotamia, specifically eastern Anatolia (now part of Turkey) and the southern Caucasus (in Soviet Georgia).

From there the protolanguage spread out and differentiated into the languages that make up this group today. The traditional picture invokes conquering horsemen imposing their language by force. But another scenario, propounded recently by archaeologist Colin Renfrew of the University of Cambridge, depicts a gradual diffusion of language borne not by the warrior's chariot but by the farmer's plow [see "The Origins of Indo-European Languages," by Colin Renfrew; SCIENTIFIC AMERICAN, October 1989].

Examining archaeological evidence, Renfrew concluded that farmers' offspring, moving just a short distance away from their birthplaces, could have spread the languages across Europe in the space of 1,500 years. As people moved out of contact, their speech changed, from generation to generation, into distinct dialects, then mutually unintelligible languages. Although it would appear that no such agricultural mechanism could have guided the dispersion and differentiation of languages spoken before the invention of agriculture, Renfrew says he is intrigued by the Nostratic hypothesis. "The more I've heard of Nostratic, the more interested I have become," he says.

The traditional linguists believe the evidence of their reconstructions is trustworthy. Reconstructions, they say, meet the strictest definitions of science: they make predictions that can be tested positively against empirical evidence. A famous example dates to the 19th century, when Ferdinand de Saussure, a French linguist, deduced that the Indo-European languages were descended from a system that included a class of sounds that had not survived in any known tongue. Later scholars identified the sounds as laryngeal consonants, so called because they are produced at the back of the throat.

Saussure's theory was regarded as a neat but rather artificial construct until the 1920s, when archaeologists dug up tablets from the archives of ancient Hittite kings at Hattusas, in modern Turkey. The tablets were found to be written in the previously unknown Anatolian languages, which preserve some of the original proto-Indo-European consonants that Saussure had predicted.

Nostraticists argue that their comparisons produce equally valid insights into far more ancient cultures. Nostratic is held to be ancestral to Indo-European, the Dravidian languages of India, the Kartvelian languages of the southern Caucasus, the Uralic family (including Finnish and Samoyed), Altaic (such as Turkish, Mongolian) and Afro-Asiatic or, as it is sometimes called, Hamito-Semitic (such as Arabic, Berber).

Illych-Svitych died at the age of 31 years in a traffic accident, leaving his projected dictionary of Nostratic roots incomplete. Dolgopolsky, who emigrated to Israel in the 1970s, carried on the work at the University of Haifa. He now has about 1,600 roots, many of which carry cultural baggage comparable to that in the reconstructed lexicon of Indo-European.

The differences, however, are much more interesting: Nostratic has many words for plants but none for cultivated varieties or for the technologies of cultivation. Similarly, it has words for animals but does not differentiate between domestic and wild. From these absences of data, scholars tentatively conclude that the language was spoken before the emergence of agriculture and animal husbandry.

It would therefore appear that speakers of Nostratic were hunter-gatherers. Vitaly Shevoroshkin, a former associate of the Soviet Nostraticists now teaching at the University of Michigan, observes that the vocabulary includes words like *haya*, which meant the running down of game over a period of days. Still, he notes that words for fairly permanent shelters indicate that the Nostrates (as he calls them) lived in villages for which some tentative archaeological evidence exists.

To arrive at those conclusions, however, the Nostraticists are accused of taking shortcuts that traditional Indo-Europeanists consider unacceptable. The traditionalists admit that the Soviet linguists indeed reconstruct the requisite web of sound shifts and try to reconstruct protowords according to accepted rules. But, they maintain, the Nostraticists compare poorly sifted data and accept as cognates words from different languages that might owe their resemblances to chance or simple linguistic borrowing.

The Nostraticists reply that they minimize chance resemblances by seeking complex sound correspondences and that they rule out borrowing by examining words that are particularly un-

A Discipline Where Caution Prevails

Linguists tend to conservatism, and not without some justification. More than most other scholars, linguists are besieged by amateurs advancing potentially embarrassing notions. Most of these amateurs are merely harmless drudges in the grip of a private theory. But there also was Adolf Hitler, who used a strange combination of Darwinism and Indo-European linguistics to support his ideology of an Aryan superrace.

With the taint of World War II, it is perhaps not surprising that many historical linguists are now quick to dismiss any attempt to find connections between genes and language. Even linguists who specialize in matters besides classification tend to be leery of any idea that might hint at racist implications.

Take the theory about Creole languages propounded nearly 10 years ago by Derek Bickerton of the University of Hawaii. Bickerton studied the common languages created by immigrants to island communities and found that they changed strikingly between the first and second generation of speakers. First-generation speakers cobble together structurally cumbersome codes called pidgins, whereas their children create complete languages called Creoles. Bickerton argued that all Creoles share structural features that reflect innate patterns of mind.

That's not the equalitarian idea it might seem to be at first, says Mark R. Hale of Harvard University. "That Creoles somehow preserve a structure innate in children under the age of two is, in a way, racist or can be used for that," he says. Might not the invariant syntax of Creoles prove, rather, the unity of humankind? "I think that is what Bickerton intended," Hale replies, after a pause. "But there is the danger that others might put another construction on it. Remember, these Creoles are spoken by brown people, the children of slaves and indentured laborers."

Even studying the words used for colors can get a linguist into political trouble. In 1968 Paul Kay and Brent Berlin of the University of California at Berkeley set out to test the belief that people perceive the world wholly through the conceptual grid of language. The extreme position held that a Russian, for example, might be conditioned by his language to see a color as light blue or dark blue, whereas an American might see it as plain old blue, for which a single English word exists.

Kay and Berlin found that it does not work that way. "Languages that lump many colors together under one color term tend to lump in the same way," Kay observes. "Some people haven't forgiven me to this day," he says. "Some even argued that my theory had racist implications—not the universalist aspect, but the calling attention to the varying number of color terms. The languages of Colonial oppressors tend to have more words for color than those of the oppressed."

Woe, too, to a biologist who links languages to population groups. Stanford University geneticist Luigi L. Cavalli-Sforza was attacked for correlating the frequency with which certain genes appear in populations to the languages those populations speak. He thinks genes and language diverged at the same time—when populations split apart—and that some splits date to the settling of the earth by modern *Homo sapiens.*

It was Charles Darwin who first linked the evolution of languages to biology. In *The Descent of Man* (1871), he wrote, "The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel." But linguists cringe at the idea that evolution might transform simple languages into complex ones.

Today it is believed that no language is, in any basic way, "prior" to any other, living or dead. Language alters even as we speak it, but it neither improves nor degenerates. Modern English may be better than Old English for a discussion of physics, but this tells us nothing about the potential resources of either tongue.

Earlier generations of linguists had no such reservations about ranking languages. In the 19th century August Schleicher classified them according to their structures.

Schleicher said that "isolating" languages, such as Chinese, used simple elements and were thus more "primitive" than "agglutinating" languages, such as Turkish, which builds its words from distinct forms. He put "inflecting" languages like German higher because their words vary with function. He ranked Sanskrit highest because its inflections were so elaborate. The system failed even on its own terms, however, by elevating the languages of many hunter-gatherers above those of the linguists themselves.

likely to be borrowed. These so-called stable words denote concepts believed to be common to all languages—body parts, for example, or natural objects such as the sun and the moon.

Lost in Static

"It is certain that the word for house is much less stable than the word for hand, because the whole system of architecture and house building can be replaced, and the words with them," says Sergei Starostin, a Soviet linguist. "It is known that you borrow things and the words that describe them. But you do not borrow *hand;* it does not happen."

That's not good enough for the Indo-Europeanists. "I'm not saying the Nostraticists are wrong. I'm saying they haven't shown that they are right," declares Eric P. Hamp of the University of Chicago, one of the world's leading historical linguists. Hamp and other conservative linguists argue that the information left over from long-dead languages must tend, like the signal broadcast by a distant radio station, to be drowned in the static of random linguistic change.

For their part, the radicals say Hamp and his like-minded colleagues take potshots at Nostratic and related theories without making a reasoned response. "I have repeatedly urged Hamp and Calvert Watkins at Harvard to publish their objections to Nostratic," says Alexis Manaster Ramer of Wayne State University. "Both of them attended a conference on these things, as discussants, in 1984 in Ann Arbor, and I repeatedly challenged both of them to state specifically what the hell they were objecting to, and they wouldn't."

In an interview, however, Hamp is voluble enough to maintain a crackling telephone conversation for six hours without a break and energetic enough to regard his imminent retirement as a chance to do still more research. He readily wrings out citations from "your schoolboy Latin" and "your Homeric Greek" without considering that his interlocutor may have little of the one and less of the other.

Nor is he reticent as he listens to the Nostratic root for dog, wolf, namely, *küjna/qüjna.* "That alternation between *k* and *q* is already unacceptable," he says. "We do that with an occasional form, but only when we think we know the phonological system." When he hears the root's proposed meaning, *dog, wolf,* he dismisses the entire exercise. "I would say they're playing very fast and loose with the semantic content," he says. "From all we know of

proto-Indo-European, the word for dog was used for the domesticated thing, which apart from its zoological classification had a social content. Somebody who says, 'We'll just move *dog* over into *wolf*' has a consummately naive view of the world."

What about in the Neolithic world of Nostratic? "If Nostratic was spoken at a time when the dog was just being domesticated, then it would be perfectly reasonable" to combine the concepts, Manaster Ramer argues. Dolgopolsky adds: "I just don't think [Hamp's] right that we aren't as accurate."

If the Nostraticists bend the rules of the game, Greenberg and his group break them. That's because Greenberg does not even bother to reconstruct roots. Instead Greenberg works with large groups of languages at once—a process he calls multilateral comparison. Compare, he says, the 25 major languages of modern Europe according to the sounds they use to signify nine basic concepts—one, two, three, head, eye, ear, nose, mouth and tooth.

Under *one*, you find most of the words falling into one class or another. The similarity between *vienas* and *viens* in Lithuanian and Latvian is striking. Under *two*, some of the foggier boundaries become clear—such as *dau, dau* and *do*, for Breton, Irish and Welsh, respectively. "By the time you get through, you've got three main groups: Indo-European, Finno-Ugric and Basque," Greenberg says. "Then you can divide Indo-European into Romance, Baltic, Slavic, Germanic and Albanian."

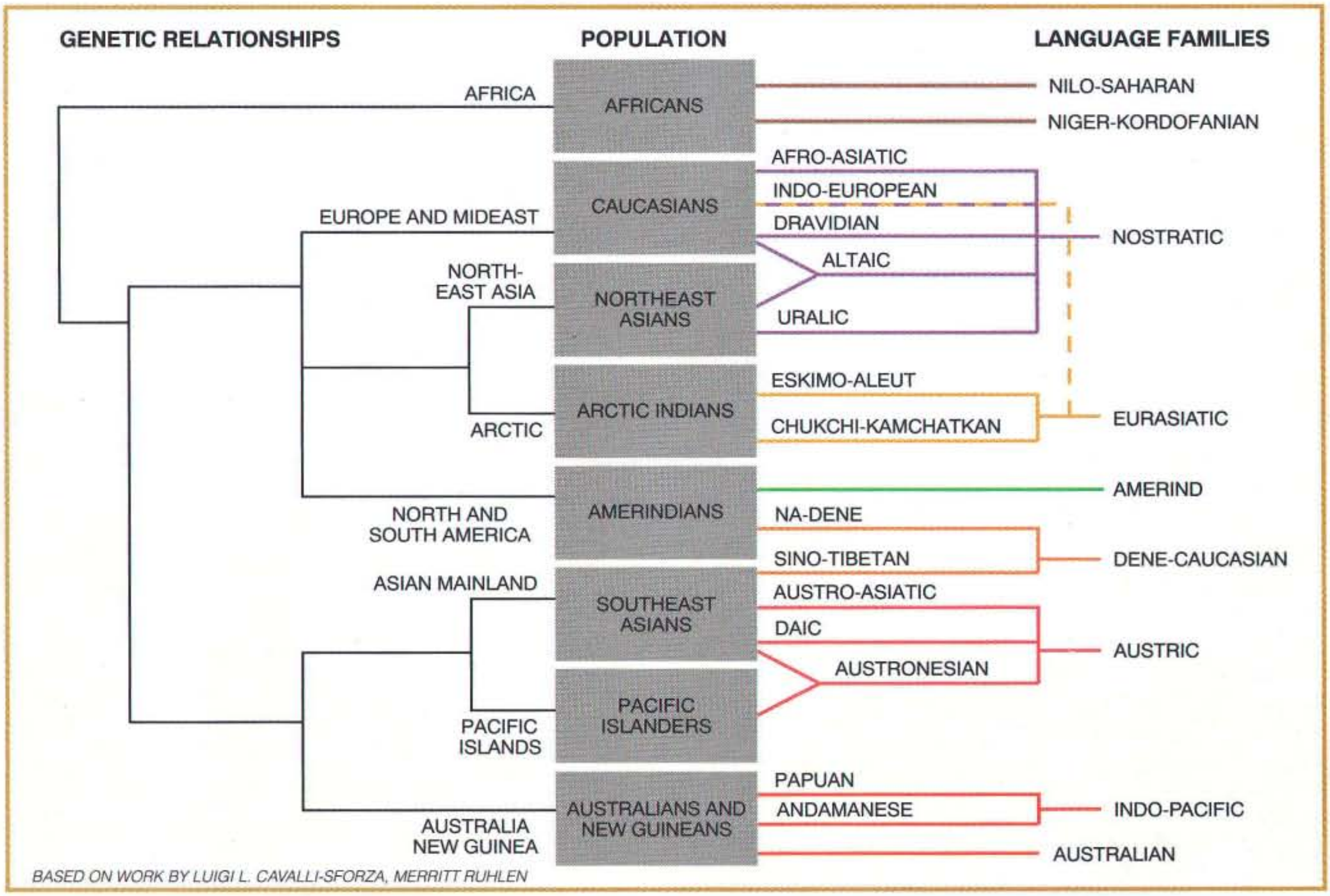
Greenberg says his multilateral comparison is so powerful that it will find relationships among languages even when the data available are rather thin. He first applied the method nearly 30 years ago to a reclassification of the African languages—many of them rather poorly documented. His work won him praise, even from some of his most ardent critics. His success may have encouraged him to enter into the greatest snake pit of his life, the notoriously diverse languages of the New World.

When he classified American languages, Greenberg divided them first into Eskimo-Aleut and Na-Dene, which has members in the Pacific Northwest and the southwestern U.S., such as Navaho. Both groups are generally accepted. But the specialists bitterly attack Greenberg's third group, which he dubbed Amerind, in which he throws the hemisphere's many other language families.

While Nostratic raises the traditionalists' hackles to half-mast, Greenberg's Amerind thesis completes the job. The reason is the apparent ease with which his system solves one of the thorniest problems in contemporary anthropology. There are 150-odd Native American language families, each apparently as different—the specialists say—as Indo-European is from Sino-Tibetan. Yet the Old World has only 40-odd families, despite the much more ancient history of settlement that ought to have produced more linguistic diversity, not less.

Moreover, crucial data vanish whenever a speaker's death puts an end to the last language of a given family, and there are not enough linguists (or funding) to describe the dying tongues. Linguists estimate that half of the world's 6,000 existing languages will die out in the next century. No one is even sure how many there might have been. "Even when you have five living speakers left, you don't see linguists rushing around with an ambulance and a tape recorder," says Thomas L. Markey, who has recently organized conferences on Indo-European and Nostratic studies.

Even without attempting to recon-



GENETIC AND LINGUISTIC histories roughly correspond because both diverge when populations split apart to form isolated communities. The genetic closeness of populations is shown by their distance from a common branching point (*left*); their linguistic closeness is depicted in a similar chart of language families and superfamilies (*right*).

Are Humans Born to Speak?

Language is the greatest human invention—if it is an invention at all. Noam Chomsky of the Massachusetts Institute of Technology thinks it isn't. He believes that language is as innate in the infant as flight is in the eaglet and that children do not so much learn language as develop it spontaneously in response to a stimulus. "Very few people are concerned with the origin of language because most consider it a hopeless question," Chomsky says.

Indeed, the question had generated so much talk and so little knowledge by 1866 that the Linguistic Society of Paris banned its discussion. The ban failed long before Chomsky came along, however, and many theories have been advanced:

- The bow-wow theory: Early words may have formed by onomatopoeia, as in *bow-wow* for dog, *cuckoo* for the familiar bird and *whoosh* for a puff of wind.
- The pooh-pooh theory: Emotional interjections such as *pooh, bah* and *harrumph* may have constituted early words.
- The yo-he-ho theory: When many people coordinated their exertions to haul a rope or roll a stone, they may have had recourse to ritualized chants that eventually acquired meaning.
- The la-la theory: Some sounds may have originated in play, as in children's singing or lovers' cooing.
- Oral gestures: Early speakers may have used their lips to point, creating vowel shifts that distinguish close from far. This may explain the shifts heard in the English words *this* and *that* and in the French *voici* (here it is) and *voilà* (there it is).

That, of course, begs the question of why language arose at all. Derek Bickerton of the University of Hawaii speculates in his recent book *Language and Species* that the development of the brain produced language as a by-product. Neural structures that allowed early hominids to abstract from their perceptions a "secondary representation" of the world improved their capacity to adapt to their environment. These structures then might have enabled them to attach meaning to gestures and sounds, producing a primitive language that had no syntax, the ordering of words that defines human language.

Primitive language would have been full of words signifying meaning but would have lacked grammatical elements. Because it would have evolved much earlier than syntax, traces of it may perhaps be found in the gesturing of apes that have been taught the elements of sign language. Bickerton speculates that feral children also provide an inkling of what primitive language was like because they were isolated in the crucial early years when syntax normally develops. Supposed "wolf children" rescued from the wilds of India early in this century could learn many things, but they never advanced much beyond the verbal ability of a normal two-year-old.

The jump from primitive language to syntactic language is hardest of all to explain. Bickerton offers a tentative argument that concludes that "a single genetic event might indeed have been enough to turn protolanguage into syntacticized language." He concludes by noting that all the prerequisites for language—larger brain, improved vocal tract, new neural linkages—involve changes in the anatomy of the head.

Chomsky, too, agrees that the event must have happened abruptly in evolutionary terms because syntax reflects an in-

struct ancient tongues, Greenberg insists that his method is illuminating the distant past. And a growing number of scientists in other disciplines agree. Two early converts were Stephen L. Zegura and Christy G. Turner II of the University of Arizona. They had studied genetic and dental variation among Native Americans independently of Greenberg. When they heard him lecture on his preliminary findings for Native American languages, they told him that his results closely matched their own. When the biological and linguistic classifications were plotted on a map, they roughly coincided. The main discrepancies came in the Native American populations of the Pacific Northwest, which apparently had a tangled history.

Early Americans

Greenberg and his two collaborators published their findings jointly in 1986 and concluded that the ancestors of Native Americans must have immigrated in at least three discrete waves over the land bridge that once connected Siberia to Alaska. They could not rule out more than three, because additional implanted languages and genes might have left no traces. Exactly which communities of Asia contributed to these waves is hard to say, although Soviet linguists, working independently, have suggested a link between Na-Dene and the languages of northern Caucasia.

Further genetic evidence in support of Greenberg's Amerind hypothesis has since been provided by Douglas C. Wallace of Emory University. "Data we published this year indicate that Greenberg's hypothesis seems correct," Wallace said in an interview in late November. "Our data show, I feel strongly, that the paleo-Indians—Amerinds—are one group."

The combination of genetic and linguistic analysis works particularly well in America, Wallace adds, because it was a blank slate. "The first immigrants who came over the land bridge did not encounter another culture," he says. "I'm not surprised that there are ambiguities in the Old World, because there have been so many mixtures."

But those mixtures can be explained, argues another important backer of Greenberg's hypothesis, geneticist Luigi L. Cavalli-Sforza of Stanford University. "Genes have no direct effect on language," he notes, "but the language that you learn depends on where you were born and with whom you were born—your family and social milieu. If a group separates, both the gene pool and the language will diverge, so the history of genes and of language is essentially the same."

Cavalli-Sforza is a venerable worker in his field, as Greenberg is in his, and he too has come under fire for supporting the Amerind theory. He went even further three years ago, when he published a study finding correlations between the frequency with which certain genes appear in populations and the families of languages that those populations speak. He took his linguistic data from Greenberg's classification of all the world's languages.

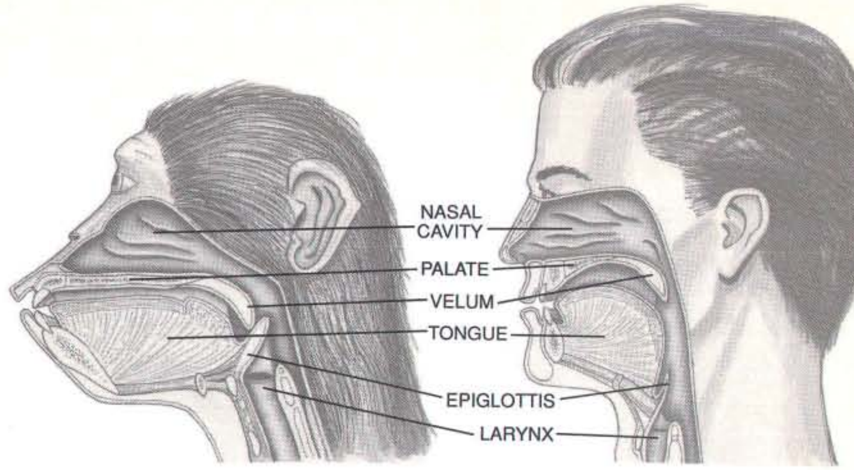
There were some notable cases where details of the linguistic and genetic trees did not match, Cavalli-Sforza says. Each, however, can be explained as the result of the replacement of one language by another, or the interchange of genes. Hungary provides a good example of language replacement: although its people are genetically like their European neighbors, they speak a language from a non-Indo-European family, adopted in the Middle Ages from Hungary's Magyar conquerors. African Americans have similarly undergone both language replacement and partial gene replacement. Not everybody accepts these explanations. "Linguists can test word resemblances to see if they result from common descent or just from borrowing, but I don't see how Cavalli-Sforza can do the same for his genetic patterns," Manaster Ramer says.

But to get a true picture of ancient demographic patterns, Cavalli-Sforza has laid particular emphasis on study-

born pattern of great intricacy. He thinks this innate ability explains why any child can learn any language without making the grammatical mistakes that one would expect if no preprogrammed structure were at hand.

Human language ability could have emerged suddenly when some genetic event knitted together a panoply of traits that had evolved for other purposes, Chomsky and others say. One such trait may have been the conscious control of vocalizations. Dogs bark when the mood seizes them; chimpanzees at least try to choke back impolitic calls, although with indifferent success. Humans, however, can lie through their teeth.

Another adaption may have been the ability to decode vocalizations. The very best telegraphists have never been able to absorb Morse code at the rate at which even a half-attentive child can extract meaning from dialogue. Our brains, it appears, are adapted to process vocal modulations.



NONHUMAN PRIMATES such as the chimpanzee (*left*) can breathe while swallowing because the epiglottis and velum form a watertight seal. In humans (*right*) the position of the larynx makes this feat impossible. Source: Philip Lieberman.

The modulations themselves have remarkable richness, in part because of the unusual shape of the human vocal tract. Philip Lieberman of Brown University notes that this shape makes humans the only mammals incapable of drinking and breathing at the same time. The drawback is that it is not uncommon for humans to choke to death while eating.

The increased risk of choking would seem to have been offset by the improved articulation made possible by our vocal tract. The Neanderthals, in contrast, had a vocal tract resembling that of an ape, according to Lieberman's reconstruction of the fossil evidence. If so, then they would have been rather inarticulate, a failing that might explain why they died out and *Homo sapiens* survived.

ing those populations that are believed to have borrowed less of their language and of their genes. "In aboriginal populations, there is no complete blurring," he says. "Traces of the common history remain, going back, as far as I can tell, to the original settling of the earth, perhaps as much as 100,000 years ago."

Mitochondrial Eve

That estimation reflects the Eve work of Allan Wilson and his colleagues. The genes they studied to trace humanity's common genetic inheritance were encoded on DNA of the mitochondria, intracellular organelles that convert glucose into a more readily used form of energy. Because such DNA is inherited strictly from the mother, without being scrambled with the father's genes during mitosis, "Mitochondrial Eve" had to be a woman.

Wilson and his team found more genetic diversity in Africa than anywhere else, so they hypothesize that Eve and her tribe lived there. They estimated the dates on the basis of two comparisons: one among humans from around the world and one between humans as a group and chimpanzees. The DNA is then used as a "molecular clock," calibrated by the estimated divergence between the human and chimpanzee lineages (which split off more than five million years ago).

Yet all the corroboration provided by the geneticists and archaeologists cannot sway Indianists such as Lyle Campbell. He rejects Greenberg's position with the vehemence of a litigator, declaring that Greenberg's theory is at once derivative and unsupported. "I don't want to give Greenberg any credit at all," Campbell snaps. The idea "was already around; we were working on it all along. This doesn't address the problem of whether these verbal similarities are due to historical reasons or to other things."

Campbell also dismisses Greenberg's word lists as worthless and his grammatical patterns as illusory. These patterns, which Greenberg regards as "conclusive," are clearest in the case of two pronominal markers, *n-* denoting the first person and *m-* denoting the second person. Campbell says that these markers are absent in many of the languages Greenberg classes as Amerind and that they are present in many he excludes. Because they are easy-to-pronounce nasals, he adds, the sounds are likely by chance to end up as grammatical markers. He also invokes the argument of "nursery talk" to link each sound to the concept it denotes.

Whether or not the Nostraticists and Greenberg are playing fast and loose with the rules, the hypothesis of the monogenesis of language is one that most linguists believe to be plausible. Indeed, the appearance of language may define modern *Homo sapiens* and explain why our species apparently did not interbreed with contemporaries like the Neanderthals. It is a horrifying scenario: the hominids without language would have seemed subhuman.

Monogenesis is not an issue that is likely to go away. The question of the origin of language is of preeminent importance: the history of language bears on the trading of goods, the migration of peoples and the evolution of ideas. Every discovery—by comparative linguists of whatever school—pushes back the borders of what is considered the irrecoverable past.

FURTHER READING

TYPOLOGY, RELATIONSHIP AND TIME. Edited and translated by Vitalij V. Shevoroshkin and T. L. Markey. Karoma Publishers, 1986.

LANGUAGE IN THE AMERICAS. Joseph H. Greenberg. Stanford University Press, 1987.

IN SEARCH OF THE INDO-EUROPEANS: LANGUAGE, ARCHAEOLOGY AND MYTH. J. P. Mallory. Thames and Hudson, 1989.

LANGUAGE AND SPECIES. Derek Bickerton. University of Chicago Press, 1990.

UNIQUELY HUMAN: THE EVOLUTION OF SPEECH, THOUGHT AND SELFLESS BEHAVIOR. Philip Lieberman. Harvard University Press, 1991.

SCIENCE AND BUSINESS

Love and Terror

Is a chemical messenger key to treating Alzheimer's?

One of the first stages of Alzheimer's disease is family denial. Acceptance comes slowly, restrained by good days that spring up between bad ones, pricked by moments when victims not only remember but are aware of not remembering. Relatives ask why, as loved ones' personalities deteriorate, then language and eventually all ability to care for oneself. Four million elderly Americans have been stricken already; the disease will claim 100,000 minds this year.

Researchers found the first solid clue connecting Alzheimer's disease and a biochemical disorder in 1976. They reported that levels of choline acetyltransferase (CAT) can be as much as 90 percent below normal in certain brain regions of Alzheimer's patients. That enzyme enables the body to build acetylcholine, a neurotransmitter critical to memory. Without CAT, acetylcholine levels (which cannot be reliably measured at autopsy) should also fall. Lack of that substance in the cerebral cortex and hippocampus, the parts of the brain where memory is thought to reside, might explain the diminished mental ability that is the hallmark of Alzheimer's disease.

They've got your number, gas antidotes, testing for integrity, retirement buy-outs

At least seven drugs based on this so-called cholinergic hypothesis are now in development. The substances are designed to protect acetylcholine from being broken down by enzymes, to cause the brain to produce more of it or to render what is there more potent.

Other scientists liken the potential effects of such drugs to bailing out a boat without fixing the leak. Because the neurons responsible for making neurotransmitters become dysfunctional as disease progresses, these workers hope to maintain and restore the cholinergic cells themselves. They are exploring treatments based on nerve growth factors and eventually hope to develop chemicals that mimic these naturally occurring proteins.

The cholinergic hypothesis for Alzheimer's disease is simply that, asserts Peter Davies, one of the first scientists to notice the depressed CAT levels. Davies, a professor of pathology and neuroscience at Albert Einstein College of Medicine, maintains, "The fact is, other things go wrong. Other transmitters are lost." Yet the effect on memory of street drugs that interfere with acetylcholine is well known, he points out: "Given the fact that certain compounds produce an Alzheimer's-like picture, it seems worthwhile to see how many symptoms one could restore by addressing this."

Some answers may come forth in March, when a Food and Drug Administration committee begins reviewing clinical test results of a controversial Alzheimer's drug called Cognex. The compound, tetrahydroaminoacridine, or tacrine, inhibits acetylcholinesterase, an enzyme that degrades acetylcholine. It was widely publicized in late 1986 after publication of a 17-patient study by William K. Summers, an affiliate of the University of California at Los Angeles, in the *New England Journal of Medicine.* This past January the FDA formally criticized the methods of that test.

To quell public clamor for the experimental drug, the National Institute on Aging in 1987 asked Warner-Lambert in Morris Plains, N.J., to conduct a large double-blind clinical trial of tacrine. Later that year the FDA participated in planning the study of 300 patients with mild to moderate-stage disease.

The study was stopped initially because of liver toxicity. But E. Peter Wolf, a spokesman for Warner-Lambert, says the problem appears to be dose related and is manageable."We think the drug is approvable," he declares.

Other acetylcholinesterase inhibitors are queuing up right behind Cognex. Hoechst-Roussel expects to seek approval for an inhibitor called velnacrine maleate, or HP-029, in early 1992. "We are encouraged by what we've seen," says Richard C. Allen, group director for neurosciences at the Somerville,

PATHOLOGIST PETER DAVIES was one of the first researchers to notice low enzyme levels in the brains of Alzheimer's patients. That observation lead to the development of drugs now in clinical trials. Photo: Kent Hanson.

SCIENCE/SCOPE®

A new hydrogen maser "atomic clock" combines a compact size suitable for space applications with the highest long-term stability ever reported for this type of device. Developed and built by Hughes Aircraft Company for the U.S. Navy, the fully automated frequency standard is about 10 times more stable than currently-used cesium beam devices. Atomic clocks use the resonance frequency of an atom to provide a precise measurement of time, but use of hydrogen maser clocks in space has been limited due to their bulkiness. Other Hughes-built atomic clocks were developed for the Defense Department's NAVSTAR Global Positioning System.

An innovative computer program dramatically reduces the hours required to model the performance of new missile designs. Called Generic Missile Simulation (GEMS), the software, created by Hughes, cuts the evaluation time of new missile designs from six months to one to 20 days, depending on the complexity of the missile. The time saving is accomplished because GEMS contains a library of generic building blocks needed for missile system simulation. These building blocks are combined, or modified, as necessary to simulate a new missile design. In the past, each new design required its own, unique simulation software.

A thermal imaging system that turns night into day for crews of U.S. Navy SH-2F Light Airborne Multi-Purpose System (LAMPS) helicopters is aiding in the fight against drugs. LAMPS helicopters, equipped with the Hughes Aircraft Company's AN/AAQ-16 Hughes Night Vision System (HNVS), have been participating in law enforcement operations in support of the Coast Guard Carribean Squadron, flying hundreds of vital law enforcement surveillance sorties, sighting and reporting many suspect surface vessels which otherwise would have gone undetected. HNVS has been installed on a variety of U.S. Army, Air Force and Navy helicopters, and a derivative of the system has been selected for the U.S. Tri-Service V-22 Osprey.

The innovative deployment of a new sonar system provides an improved means of detecting, identifying, and tracking of ocean targets. The Surveillance Towed Array Sonar Segment (SURTASS), developed by Hughes for the U.S. Navy, allows antisubmarine warfare commanders to have capabilities never before possible for the collecting and processing of undersea acoustic data. The system consists of a long line of sonar arrays towed behind a noncombatant craft. Target data is transmitted through a satellite link to land-based centers where operators can review the data on a detailed display.

Hughes Aircraft Company's Ground Systems Group and new subsidiary Hughes Aircraft Company of Canada Ltd. are looking for ATC Specialists, Systems Engineers, Systems Engineers/Proposal Managers, and Air Traffic Controllers. We're applying our creative expertise and airspace management experience to many exciting international Air Traffic Control programs, including the Canadian Automated International Air Traffic System (CAATS) and Germany's Karlsruhe Workstation Control (KATC); and there's new business on the horizon. For immediate consideration, send resume to: Bill Campbell, Hughes Aircraft Company, Ground Systems Group, Dept. S3, P.O. Box 4275, Fullerton, CA 92634. Proof of U.S. citizenship may be required. Equal opportunity employer.

For more information write to: P.O. Box 45068, Los Angeles, CA 90045-0068 USA

HUGHES

Subsidiary of GM Hughes Electronics

N.J., drug company. "It's not holy water, but it does have an effect in some patients." The firm is now trying to determine which patients benefit and to quantify their responses. Similarly, Forest Laboratories, Inc., in New York City has an acetylcholinesterase inhibitor, called Synapton, in the final stage of clinical testing. It is a controlled-release form of physostigmine.

Rather than trying to keep acetylcholine levels up by preventing enzymatic breakdown, others are trying to find drugs that cause the brain to produce more of the substance. Du Pont's experimental drug, DuP 996, stimulates release of acetylcholine, as well as two other important transmitters—dopamine and serotonin. David S. Block, development marketing manager for central nervous system diseases, will release little else about the drug, which is in clinical testing at several sites. Block says the drug will be tried in all stages of the disease but is "more likely to work in earlier stages, when neurons are most intact."

Hoechst too has a compound, dubbed HP-749, which seems to amplify acetylcholine's signal, so that even a reduced chemical message is received by nerve cells. In animals, it acts on adrenergic cells that have been linked to depression and loss of secondary memory when they deteriorate in humans. If the drug performs the same way when tests on Alzheimer's patients begin this summer, Allen says "we're hoping for a broad response in the patient population."

An alternative approach to stimulating nerve impulses involves agonists—drugs that bind to receptors in place of other substances (such as acetylcholine) and trick cells into responding. Warner-Lambert has such a drug, called CI 979, in preclinical trials. The trouble comes in trying to control the agonist's signal that turns a cell on. Constant stimulation would hopelessly jam cell circuits.

Inhibitors and agonists are sometimes referred to as "software" therapeutics because they will work only as long as the brain's hardware is intact. Fixing these functional parts—rather than their products—is a goal some scientists believe will be achieved with nerve growth factors. "We are on the verge of manipulative possibilities where we can deal with structural events in the brain," says Franz F. Hefti, a neurobiologist at the University of Southern California.

Studies in Hefti's laboratory indicate that brain-derived neurotrophic factor strengthened the function of cholinergic neurons and other neurons that make dopamine, the transmitter lacking in Parkinson's disease. Other growth factors—including the most studied one known simply as nerve growth factor—also show stimulatory effects.

So far research on nerve growth factors is still at a very early stage. Although they have the potential to be highly beneficial, these potent substances could also prove quite dangerous. They could, for example, stimulate wild connections throughout the brain that might speed loss of mental ability or disturb nerve systems outside the brain. Nor will delivering nerve growth factors be easy. Because these proteins cannot pass the blood-brain barrier, they would have to be administered via an implant directly into the brain.

Despite the drawbacks, biotechnology firms such as Genentech, Synergen and Chiron, as well as pharmaceutical companies such as Syntex and Warner-Lambert, are investigating the proteins. Researchers are also screening natural and man-made chemicals for compounds that mimic growth factors but might be taken orally or by injection.

"We need strong monitoring of Alzheimer's drugs," declares Jean Marks, associate executive director of the New York City chapter of the Alzheimer's Association. Both patients and caregivers need the protection of a rigorous scientific process, Marks says: "The pressure for an immediate fix comes out of love and terror." *—Deborah Erickson*

A Plague of Plaques

Pathologists diagnose Alzheimer's disease by looking for plaques of abnormal proteins choking a deceased patient's brain cells. To some researchers, these so-called amyloid deposits also loom as a potential root cause of the disease. That idea may have been strengthened by a recent report in *Nature,* which identified a specific abnormality in a gene for the amyloid protein.

If the work of John A. Hardy and his colleagues at St. Mary's Hospital Medical School in London stands up, "then it will unequivocally show that abnormalities in amyloid processing can be the start of the disease," says Dennis J. Selkoe, a neurologist at Harvard Medical School. Selkoe is also a consultant to Athena Neurosciences, Inc., in South San Francisco, one of several companies attempting to prevent amyloid deposition in most patients.

Hardy's team has found a mutation of the amyloid precursor protein (APP) only in a few people with an inherited form of the disease. The mutation changes just a single amino acid in the long APP molecule, but that difference seems to be unique to some Alzheimer's patients. Most cases of Alzheimer's disease do not yet seem to have a clear genetic cause.

Hardy and his co-workers were studying a family in which the early onset of Alzheimer's disease was closely linked to the inheritance of the APP gene. "Almost immediately we found a mutation," Hardy explains. A second disease-prone family had the mutation as well, although most such families that Hardy's group screened did not. The researchers also failed to find it in any of 100 normal individuals. "This suggests that the mutation might cause Alzheimer's, but it's only a suggestion," Hardy cautions. "If we find this mutation in anybody who is not at risk of getting Alzheimer's, that's disproof."

Rudy E. Tanzi, a neurogeneticist at Massachusetts General Hospital, has conducted similar surveys seeking Hardy's mutation in other patients with familial Alzheimer's disease. So far he has not found it, but he and his colleagues are continuing to look for it or any other genetic differences.

No one yet fully understands how changes in amyloid processing could cause all the other biochemical and cognitive abnormalities of Alzheimer's. Enzymes called proteases routinely carve APP into shorter peptides. In Alzheimer's patients, however, large amounts of APP are split in an atypical way, and small beta fragments are liberated. It is the beta fragments that aggregate into the senile plaques seen in the disease.

If amyloid processing is really the key to Alzheimer's, and if the release of beta proteins can be stopped, then it might be possible to thwart its development. Athena Neurosciences, working with Eli Lilly, and Cephalon, Inc., collaborating with Schering-Plough, are both pursuing that strategy, among others. According to Frank Baldino, Jr., president of Cephalon, his company has identified proteases that seem to cleave APP, and it is now developing molecules to inhibit them. *—John Rennie*

Call and Tell

Dialing 1-800 gives marketers a lot of personal information

The zip code tells the U.S. Postal Service where to deliver the mail. It also tells direct marketers what to deliver. Combining the zip code with census and other data provides marketers with a rich vein of demographic information concerning your income, buying habits and socioeconomic preference for squash instead of handball.

If all this is not enough, the past decade has given direct marketers another wedge into the collective psyche of American consumers: your telephone number. Combining the resources of massive computer data bases with the ability of an emerging "smart" telephone network to identify callers, the direct-marketing industry is using the telephone number to track down a person's name, address—and life-style. If your household is deemed "desirable" to a marketer—perhaps one of the "Pools & Patios" crowd, as one telemarketer puts it—an 800 or 900 line service representative may know it before the call is answered.

Target direct marketing is not new. A company that subscribes to an 800 or 900 service can receive a monthly listing of the numbers of callers, which can then be matched with names and addresses using a reverse telephone directory. Correlating that information with demographic data produces valuable mailing or phone lists. (An 800 call is toll free, whereas the caller pays for dialing a 900 number. A caller interested enough to pay a fee is more likely to buy a product, marketers reason.)

To the consumer, all this means that products can be more closely matched to personal tastes, with the result that the junk mail might just contain something worth buying. What's new is that information-age marketers have begun to acquire the technology to carry out this screening process instantly and without the caller's knowledge.

Beginning this year, Telesphere Communications, Inc., an Oakbrook Terrace, Ill., company with $550 million in annual sales, will offer a service to 900 subscribers that can peg the location of an incoming call using an area code and the number's three-digit prefix. Knowing where the call originates allows a salesperson to prepare a pitch. Later a reverse directory can be used to identify the caller, and a data base can determine which of 40 demographic "clusters" fits that person. In the near future, these services may be provided while the caller is still on the line.

Telesphere gets its demographic information from PRIZM, a data base owned by Claritas Corporation in Alexandria, Va. PRIZM can pinpoint a neighborhood for virtually everyone in the U.S. using census and other public demographic information. "It works on the theory that birds of a feather flock

Is Morse Code Signing off?

Dit Dit Dit. Dah Dah Dah. Dit Dit Dit. Help! Is SOS about to become 1010011 1001111 1010011? Is Morse code, that hand-me-down from the days of telegraphy, destined to become just a binary blob?

That question might well be put to the Federal Communications Commission, which, as of February 14, allowed a new type of ham radio license that does not require a Morse code test. (A code requirement still exists for licenses that permit shortwave transmissions, that is, those below 30 megahertz, which are used for around-the-world communications.)

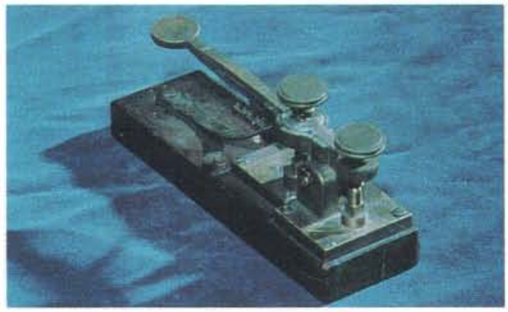
The advent of the code-free Technician Class may have inspired some purists to key out "What hath God wrought!"—the message that Samuel F. B. Morse tapped out on the world's first telegraph link from the Capitol Building in Washington, D.C., to Baltimore in May of 1844.

A code-free license had been resisted by the nation's half a million hams for years, in part because it might destroy the fraternitylike quality of owning a license. The Morse code test was also viewed as a measure of commitment to the numerous rules of operation of amateur radio—it can take weeks of study before an examinee is capable of sending the minimum of five words a minute.

The last time the FCC asked for public comment about establishing a new class of license, in 1983, the agency ended up dropping the proposal after it was opposed 25 to 1. "There has been a concern that amateur traditions might be sacrificed," says David Sumner, executive vice president of the American Radio Relay League (ARRL), the nation's largest ham organization, with 160,000 members.

It was the ARRL, among others, that proposed the present change in 1989, largely because the growing popularity of computer communications equipment is rapidly turning the amateur-radio bands into a network of wireless electronic mail. Tens of thousands of new users of what is called packet radio are transmitting computer data each year via the airwaves. The radio spectrum has also become more crowded. "There was a fear that unless we demonstrated greater receptivity to newcomers, some of our frequencies would be reallocated for commercial use," Sumner says.

Packet-radio devotees, called packeteers, can now sign on using the new license. And by sending a computer message through one of the amateur-owned satellites, they can communicate worldwide without violating the restriction on using shortwave frequencies. Amateurs have also installed 10,000 radio-repeater stations throughout the U.S. that allow a very high frequency transmission to hop from one repeater to another.



MORSE'S FIRST telegraph key. Source: The Granger Collection.

Morse code—which was actually devised by Morse's associate Alfred L. Vail—will still be used on the shortwave bands to get a message through when voice transmission would be unintelligible. Then there is the sheer pleasure that hams derive from communicating with their key sets. Sumner likens it to sailing. "People just get more satisfaction hoisting a sail rather than puttering around with the motor turned on," he says. *—Gary Stix*

together," says Harvey B. Uelk, a Telesphere sales director.

So if you are lucky, the pitchman will know if you fall in the fifth cluster in the data base: "Furs & Station Wagons." This group is described as " 'new money' living in expensive new neighborhoods.... They are winners—big producers, and big spenders." A not so fortunate caller might be lumped into the "Emergent Minorities" cluster. These people, says a promotional report, are "almost 80 percent black, the remainder largely composed of Hispanics and other foreign-born minorities.... Emergent Minorities shows...below-average levels of education and [below-average] white-collar employment. The struggle for emergence from poverty is still evident in these neighborhoods."

The risk that a household, through clustering, might become the telemarketing equivalent of a bad credit risk has not escaped the notice of the American Civil Liberties Union and other public interest groups who fear that minorities might be excluded from mortgage and credit opportunities or a gay neighborhood may be blacklisted by an insurance advertising campaign. A telemarketer might display different sales pitches on a service representative's computer screen, depending on whether the incoming caller hails from the "Money & Brains" or the "Coalburg & Corntown" cluster.

Marc Rotenberg of Computer Professionals for Social Responsibility likens calling an 800 or 900 number to walking into a store. "A person should have a right to enter a store without disclosing creditworthiness, residence or annual income," Rotenberg asserts. Lobbying by privacy groups has focused so far on supporting national legislation that would, in effect, allow a caller to keep his wallet in his back pocket until he decides to make a purchase.

The law would give the caller the option of blocking a number from being displayed immediately by a receiving party. This would be done by pressing "*-6-7," or a similar combination of numbers, before making a call. (Marketers could still get callers' 800 or 900 numbers with their statements each month, however.) Although the law failed to pass Congress last year, it is scheduled to be reintroduced this year.

Individual states are not necessarily waiting for Congress. A Pennsylvania court has banned "Caller ID" service—a decision that is on appeal—and a number of state public utility commissions have ordered that blocking be offered free of charge. For the moment, states' actions may not affect most telemarketers, whose 800 and 900 calls are usually routed over the long-distance phone network and displayed to a clerk using a service called automatic number identification.

Support for blocking has come not just from privacy advocates but from the White House's Office of Consumer Affairs, four of the seven regional Bell companies and the Direct Marketing Association in New York City. As with junk mail, the direct-marketing industry acknowledges that the consumer should have the right to choose not to receive unsolicited information.

On the opposite end of the line, a number of telephone companies contend that caller identification services are a clear boon to subscribers. Bell Atlantic, an ardent opponent of call blocking, has compiled a list of subscribers who have used the Caller ID service to stop obscene phone calls or fake pizza orders and to track down burglars.

For their part, some direct marketers assert that fears of misappropriation of personal information are greatly exaggerated: they are interested in patterns of group behavior, not the personal preferences of the individual. "We try to identify market segments that are most likely to respond to a particular marketer's products or services," explains Philip H. Bonello, director of corporate planning for Metromail, a Lombard, Ill., firm that owns a data base of 86 million households that supplies the direct-marketing industry.

But the public is clearly concerned about electronic privacy. In January Lotus Development Corporation, a Cambridge, Mass., software company, and Equifax, Inc., an Atlanta-based credit bureau, withdrew plans to market Lotus Marketplace on compact discs after some 30,000 people asked that their names be removed from the files. This data base contains demographic information on about 120 million individuals.

The public debate over privacy could grow still more heated if telephone companies try to market their internal data bases of information about residential customers. Limited attempts to do so have sometimes met with resistance. Recently New England Telephone and New York Telephone dropped a service offering residential and business directory listings when hundreds of thousands of customers asked that their names be taken off the lists.

Legislation may help stem abuses. A public outcry may force companies to lay low. But the irresistible lure of knowing name, phone number and lifestyle means that computerized telemarketing is here to stay. *Caveat salutator:* let the caller beware. *—Gary Stix*

Gas Vaccine

Bioengineered immunization could shield against nerve gas

A year ago the Defense Logistics Agency, an arm of the Department of Defense, had to fight to keep a budget item for a $2.5-million contract with Abbott Laboratories in Chicago to maintain its sole supply of atropine, a nerve gas antidote. Times have changed. Since last August the U.S. Army has ordered more than a billion atropine injectors at a total cost of $4.3 million and has had to look for additional suppliers.

Even when fully stockpiled, however, existing treatments for nerve toxins, organophosphorous compounds that pose the largest chemical battle and terrorist threat, are awkward to use and often produce side effects. So the army has also speeded up research efforts to make a genetically engineered immunization against nerve agents. If it works, the injection would free soldiers from debilitating garb that now limits movement and breathing, and it could reduce the efficacy of chemical weapons. The "gas vaccine" will not be available for years. But the renewed threat from chemical warfare will likely persist long after the conflict in the Persian Gulf ends.

For now, though, soldiers in the Gulf will continue to carry their "Mark I" kits with six autoinjectors—three with two milligrams of atropine citrate and three with 600 milligrams of pralidoxime chloride. The first compound binds with a nerve agent, preventing it from attaching to acetylcholinesterase, an enzyme that controls the transmission of nerve impulses.

The second drug frees enzyme molecules that have already been bound by the toxin. Other elements of the battlefield pharmacopoeia are an anticonvulsant that is a close relative of Valium and pretreatment tablets of a drug for myasthenia gravis, a disease that causes neuromuscular weakness.

Army officials say that these kits adequately protect U.S. troops. Still, current antidotes require that soldiers stop in the heat of battle after recognizing the symptoms of a nerve agent. They must administer the antidote within two minutes or else risk absorbing a lethal dose. The protective drugs may also slightly impair fine motor skills, affecting split-second judgment needed in combat.

Although researchers have tried to devise better means of delivery, such as time-release drugs, dermal patches

and topical ointments, what is really needed is a way to eliminate antidotes altogether. This protection "would be head and shoulders above what's available now," says Lt. Col. Richard P. Solana, chief of pharmacology at the U.S. Army Medical Research Institute of Chemical Defense at Aberdeen Proving Ground, Md.

The simplest solution would be to inject extra acetylcholinesterase or butyrylcholinesterase, a related enzyme, to bind to a toxin. Last September Bio-Technology General Corporation in New York City received a three-year contract from the Army Medical Command to engineer acetylcholinesterase genetically. One drawback: the enzyme is a large molecule that would require massive injections to be effective.

A better approach would be a genetically engineered vaccine that would circulate continuously in the bloodstream, waiting to "scavenge" a nerve toxin before it reached acetylcholinesterase. Developing a gene-spliced antibody that can be injected in reasonable doses is part of a program supervised by Solana at Aberdeen Proving Ground.

Although the army declined to supply specific information about its work, researchers there are looking for a long-lived antibody that also demonstrates reverse-binding properties, the ability to neutralize one molecule of toxin before moving onto the next. The ultimate goal, which is years away, would be to produce copies of the receptor on acetylcholinesterase that binds to nerve toxins.

While this work continues, the army's Chemical Research Development and Engineering Center at the Edgewood area of Aberdeen Proving Ground—along with a number of universities—is trying to create better field detectors that would sense the most minute presence of noxious chemicals, giving troops more time to don protective gear. Hand-held detectors and chemically sensitive paper attached to garments can sense the presence only of known nerve agents.

James J. Valdes, a neuropharmacologist at the center, heads a multimillion-dollar, decade-long research program to use bioengineering methods that would detect virtually any chemical or biological toxin, including any that an adversary may have newly concocted.

Most work involves placing a human enzyme on a gel base in contact with an electronic transducer. When a toxin binds to the enzyme, it produces a change in the enzyme's capacitance that can be detected by the transducer. Prototypes that can detect femtomolar concentrations of toxins within seconds have been assembled, but volume production of biosensors for field usage is still five to eight years away because of the trouble in creating synthetic enzymes capable of staying stable for a prolonged period. Better transducers that can survive the abuse of the battlefield must also be devised.

NERVE GAS ANTIDOTES are carried by soldiers in the Persian Gulf. Source: Tannenbaum/Sygma.

Despite the difficulties, experts at Aberdeen on this most inhumane form of warfare remain optimistic. "The army laughed in 1985 when we started this work," Valdes says. "They told us the only biosensor we'd ever have would be 100 dead soldiers and a medic to examine them. Now practical hardware is a real possibility." —*Mark Fischetti*

Mind Reader

Do personality tests pick out bad apples?

For years, employers have given job applicants paper-and-pencil tests to assess basic skills such as reading and arithmetic. These days, candidates may be confronted with a different type of exam as well: a personality test that asks about their attitudes toward a variety of situations. Answers are often run through a computer to produce a profile that rates the applicant on scales, or "personality constructs," as they are called in the jargon of the trade, with names like "gregarious" and "tough-minded."

But widespread screening worries some psychologists and personality researchers. The tests, marketed in the U.S. by such companies as the Institute for Personality and Ability Testing in Champaign, Ill., Consulting Psychologists Press in Palo Alto, Calif., and the Psychological Corporation in San Antonio, Tex., are being used by employers to exclude those not suited for sensitive jobs, in police departments and nuclear power plants, for example. They are also a means of selecting candidates with desirable traits, such as extroversion, for marketing positions.

Rodney L. Lowman, a psychologist at Duke University Medical Center and author of a book about testing, offers some words of caution: "There is far more practice than there is research literature to support the proactive use of these tests." Lowman believes that "mistakes are being made by screening services that may be overly aggressive at weeding people out." And the desirable characteristics used for screening applicants "often read like a Boy Scout list of virtues, and the specific job relevancy has yet to be demonstrated." Lowman points out that one commonly used test—the Minnesota Multiphasic Personality Inventory, published by the University of Minnesota Press—was designed for clinical patients rather than job applicants.

In addition, few commercial personality tests have been validated in published studies. The confirming studies that have appeared, charge Steve Blinkhorn and Charles Johnson, industrial psychologists at Psychometric Research and Development Ltd. in England, are so full of statistical errors that it is doubtful whether most of the constructs predict anything. The two fired a broadside at preemployment personality testing in the December 20–27, 1990, issue of *Nature,* where

Are You Honest?

Some Questions Asked on Preemployment Integrity Tests

- How often do you tell the truth?
- How often do you make your bed?
- Do you think you are too honest to take something that is not yours?
- True or false: I like to create excitement.
- How much do you dislike doing what someone tells you to do?
- Do you feel guilty when you do something you should not do?
- True or false: A lot of times I went against my parents' wishes.
- Do you believe that taking paper or pens without permission from a place where you work is stealing?
- What percentage of the people you know are so honest they wouldn't steal at all?
- How often do you blush?
- How many people have cheated on their income tax returns?
- How easy is it to get away with stealing?
- On the average, how often during the week do you go to parties?

SOURCE: Office of Technology Assessment

they accused psychologists of adopting "an approach to correlation that would have left its inventor Karl Pearson gasping."

Usually, the British critics suggest, developers of tests have identified apparently significant associations by sifting through thousands of combinations of response patterns and job performance measures. But such trawls throw up spurious correlations by blind chance, a complication that test marketers tend to overlook. Nor have the test services often checked their claimed correlations against new data. Blinkhorn and Johnson found that for three well-known commercial tests, most of the supposed correlations between scores and job performance were likely to be the result of pure chance.

Other critics are concerned about tests that purport to measure honesty or integrity, which have become increasingly popular since polygraph testing by businesses was banned in 1988. More than five million people in the U.S. take honesty tests each year, according to one estimate. These tests attempt to flag job candidates who are likely to steal property or company time by asking about their past behavior and their attitudes to various types of theft.

Despite the obvious possibilities for cheating, some sellers claim the tests can predict subsequent inventory disappearances or detect previous criminal behavior. But validating such tests is difficult because employee-thieves are seldom caught. In addition, publishers of honesty tests "intentionally err on the side of lenience" to avoid making false accusations, contends Richard E. Clingenpeel, who runs Personnel Selection International, a job agency in Milford, Mich.

According to a draft report on honesty tests by the American Psychological Association, "a few firms have made public a number of reports of studies having to do with the reliability and validity of their instruments, but most firms have produced nothing of this sort." The congressional Office of Technology Assessment concluded recently that published validations of integrity tests are flawed and inadequate.

Still, basic personality tests have their proponents, who argue that they can be a useful screening tool. "I agree there's plenty of very bad research using personality tests, but I wouldn't characterize the whole field that way," says Paul R. Sackett, a psychologist at the University of Minnesota.

Since the mid-1980s a consensus has emerged that there are five or six robust personality dimensions out of the many more that have been proposed. Michael K. Mount, an organizational psychologist at the University of Iowa, groups personality constructs into one or other of what he calls the Big Five categories (extroversion, emotional stability, agreeableness, conscientiousness and openness to experience).

Mount and some other psychologists say they are finding modest correlations between such personality traits and job performance. Using the emerging technique of meta-analysis, which rigorously compares data collected in different studies, Mount and his colleague Murray R. Barrick surveyed 117 studies of personality traits and job performance for a paper that will be published shortly in *Personnel Psychology*. "In a nutshell, most personality indicators were not good predictors," Mount says. One measure, conscientiousness, did show a persistent correlation with job performance. But the correlation is about half as strong as that achieved by mental ability tests.

The U.S. Army has for some years been conducting an exercise, known cryptically as Project A, to develop personality tests. Long-term validation studies with large numbers of subjects are under way. Leaetta M. Hough of the Personnel Decisions Research Institute in Minneapolis, the principal contractor for Project A, has also used meta-analysis to show that some aggregated personality constructs may predict something about job performance. One of these, called dependability, is similar to Mount's conscientiousness construct.

Ivan Robertson, a psychologist at the University of Manchester in England, agrees that meta-analysis shows that "there's a definite effect of personality—but it's small." Properly constructed personality tests seem to give unique information that could be useful to employers if used with other forms of screening, Robertson says.

So will meta-analysis bring validated personality tests that are the answer to a corporate recruiter's prayer? Blinkhorn doubts it because of the difficulty in training testers. Moreover, he points out that if Hough's results are typical, a test would have to be used very stringently to hire noticeably better workers. If only the best-scoring 10 percent of candidates were hired, the proportion of above-average workers in a group would rise from 50 percent to only 59 percent, and many good workers would be wrongly rejected.

The suspicion that personality tests unjustly reject some applicants is probably why they are unpopular, admits Scott Martin, a psychologist at London House, a test publisher in Park Ridge, Ill. "Once you make it objective, it gives people something to criticize." But Martin argues that even a poor test will erroneously turn away fewer good candidates than selecting at random.

Martin's argument may be irrelevant in the real world. Although some preemployment test publishers say their tests should accompany other forms of assessment, such as an interview, in practice personality and honesty tests are often used instead of other forms of assessment. So they could be a step backward. Lou Maltby of the American Civil Liberties Union fears that tests may create an unemployable group who "test dishonest" or otherwise prove unsuitable. If Maltby is right, pseudopsychology will hurt employers as well as employees. —*Tim Beardsley*

THE ANALYTICAL ECONOMIST

Golden handshakes, golden handcuffs

When times get bad, employers often cut costs by offering early retirement. Losing older, more highly paid workers saves money regardless of whether job cuts are permanent or whether younger, lower-paid replacements are eventually hired. Although economists typically discount the notion that grizzled veterans may offer such unquantifiable extras as corporate memory, such retirement plans have implications beyond short-term cost cutting.

Early-retirement incentives bring into sharp focus the economic underpinnings of the so-called social contract between employer and employee. Now that mandatory retirement has been largely abolished in the U.S., all retirement is "early." The experience employers have had with early retirement thus far may presage the behavior of a generation of employees now reaching middle age with the knowledge that no set date delimits their working life.

According to labor economists, companies traditionally overpay older employees and underpay younger ones in the same job. Some see this compensation scheme as a form of worker discipline: anyone who is fired at an early age will miss out on well-paid dotage. Others count the overpayment of older employees as deferred wages. And a third camp considers taking care of older employees a part of the cost firms must pay for high-quality workers.

This implicit social contract has been formalized at least in part. Laws forbid discrimination against older workers, and many contracts protect them from layoffs. James L. Medoff of Harvard University and Katharine G. Abraham of the University of Maryland have found that in more than 80 percent of firms, older workers either are never laid off before younger ones or are laid off first only if a younger worker is considered significantly more valuable.

Nevertheless, companies may feel compelled to break this social contract, particularly during a recession. Any company that can shape its workforce—without either breaking the law or destroying employee morale—into one with a larger than average proportion of younger, lower-paid workers will tend to be more profitable. Not only will overall wages be lower, but jobs may get done better as well. Medoff and Abraham found that workers who have been doing the same job for many years may be less productive than newcomers. Long tenure in the same position, Medoff says, implies that an employee has been consistently passed over for advancement.

Offering older workers incentives to retire early is one way to slip out of the social contract. Essentially the company splits the benefits of having a younger workforce with those older employees who step aside to make way. The buy-out could be an extra year's pay, spread over two years, special pension benefits—such as IBM's recent waiver of a minimum age for full pension—or even a lump-sum settlement.

But early-retirement offers can backfire. Sometimes not enough people leave; sometimes, too many. In 1984 New York State found itself hiring nearly 4,000 people to replace those who jumped ship. And in 1985 Du Pont found itself severely short-staffed when 11,500 employees, instead of the 5,000 expected, accepted its generous retirement terms. "There's an awful lot of crapshooting going on out there," says Olivia S. Mitchell, a labor economist at Cornell University.

Even when the numbers are right, retirement incentives can still have a perverse effect: the best performers leave, and the worst ones stay. This makes perfect economic sense, Mitchell says: talented workers can find comparable jobs somewhere else and still collect a pension. Mediocre people are less likely to succeed at double-dipping.

To protect themselves from such defections, some employers insert a "business need" clause in the early-retirement offer that permits them to veto the departure of vital employees. But invoking that clause may lead to litigation. Others tilt the value of the buyout: Stanford University, for example, has an early-retirement incentive designed so that those valued least by their departments get the most when they leave. Yet there are no good data on whether such targeted incentives achieve their purpose, says Albert Rees of Princeton University, who specializes in academic retirement.

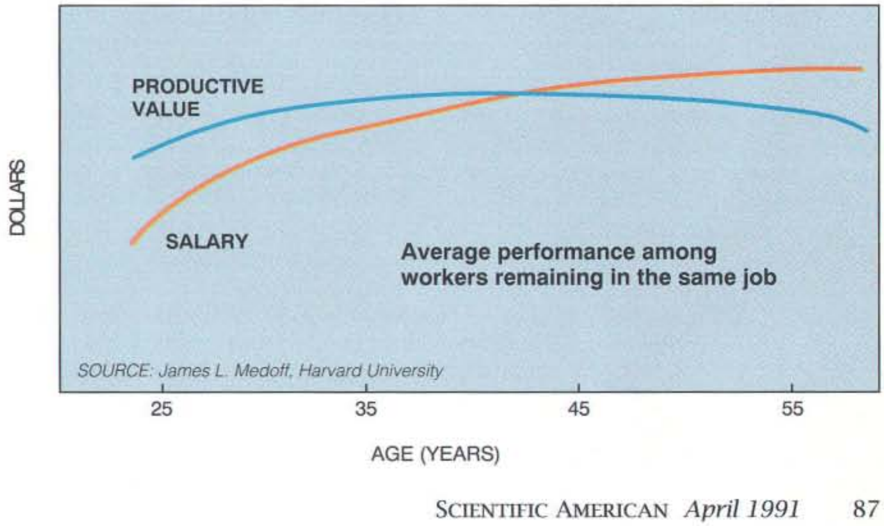
It is consequently unclear how such incentives might function if firms have to buy out almost all older workers. Indeed, it is unclear how the social contract of employment itself might be modified if there is no predictable end to the period over which older workers may be overpaid. Legal issues aside, wholesale abandonment of the old patterns could leave workers feeling betrayed and with little financial incentive for corporate loyalty.

Partial retirement offers one possible out. Alan L. Gustman of Dartmouth College observes that about a third of all workers currently ease out of the workforce by retiring from their primary job and then finding a less demanding job elsewhere. Fast food and hotel chains are among those who have found it advantageous to hire older workers on a part-time basis. Gustman suggests that even more people would take this option if most companies did not still insist that employees work nearly full time or not at all.

Partial retirement could offer a workable employment alternative for older workers willing to accept less pay for less work. But what about those relatively overpaid employees who want to keep on working full time? The good news is that there may not be many. "There just aren't that many people who want to work past 70," Gustman comments. Thus far the social contract seems to be holding up.

—Paul Wallich and Elizabeth Corcoran

How Wages and Performance Vary with Age



MATHEMATICAL RECREATIONS

Why Tarzan and Jane can walk in step with the animals that roam the jungle

by Ian Stewart

Tarzan leaped into the air, kicked both legs out in front and fell abruptly on his derriere. Jane watched from the tree house as he repeated the maneuver 20 times. She began to doubt whether her muscle-bound companion had any brains at all. She had spent months trying to teach him basic science and English, but he had resisted any ideas that were contrary to the laws of the jungle. Grabbing a nearby vine, she swung down to find out why he was monkeying around.

"Good Morning, my radiant angel," cheered the ape-man.

She was glad to note that his vocabulary was improving. "Were you jumping around for some reason, or were you just massaging your buttocks?"

"I was testing Curie's principle."

It was a novel excuse, she thought.

"And I think it doesn't work."

Jane took his hand. "Let's go somewhere cool and quiet, and you can tell me all about it."

Under a shady tree at the end of a field, Tarzan began to explain. "I recently read that the human body is bilaterally symmetric. Which means that it looks pretty much the same when reflected in the surface of a still pond or some other mirror. Right?"

Jane nodded in agreement.

"Then I came across a fundamental principle proposed by the physicist Pierre Curie. He claimed that symmetric causes produce equally symmetric effects. So it seemed to me that if I'm a bilaterally symmetric man, I should walk in a bilaterally symmetric way. To accomplish this, I must kick both legs forward at the same time."

Tarzan demonstrated the motion for Jane one more time. Falling on a sharp stick, he cried out, "*Aayiyiyaah!*"

"You've missed the point," Jane said.

"No, I haven't," Tarzan countered, pointing to the sizable mark on his leg.

"I mean you misinterpreted Curie's principle. First of all, if you want a bilaterally symmetric gait, you can hop like a kangaroo." She held her hands near her chest and leaped several times with her feet together. Tarzan watched the spectacle, fascinated. Finally, he mustered enough courage to ask what a gait was.

"It's a pattern of limb movement used for locomotion," Jane explained. "Animals use all kinds of different gaits to get around. Humans walk, horses gallop. Gazelles even pronk, that is, they hop on all four legs."

"Hopping is all very well," said Tarzan, "but all it shows is that a symmetric gait is possible. My reading of Curie's principle is that all human gaits—in fact, all gaits of bilaterally symmetric animals—ought to be bilaterally symmetric." He paced thoughtfully up and down the clearing. "But most of them aren't."

Jane watched, trying to imagine what his walk would look like in a mirror. "It almost is. When you reflect a walking gait, it still looks like a walking gait." She paused. "It has to, really, otherwise people walking would look funny in a mirror. Although I suppose that's not conclusive."

"The difference," commented Tarzan, "is that when I put my right foot forward, my mirror image puts its left foot forward. Now, on my next step, I do put my left foot forward, but by then my mirror image is putting its right foot forward. We're always out of step with each other."

"Out of phase, not step," Jane corrected. "That's why everything looks all right in a mirror. If you delay time by the amount required to take one step, then the mirror walk looks exactly the same as the original."

"Phase?"

"Walking—like all gaits—is a periodic motion. It repeats at regular intervals of time. If you have two copies of the same periodic motion and one is time-delayed relative to the other, then the fraction of the period representing the delay is called the relative phase. Your left leg is out of phase with your right leg by exactly half a period, that is, a relative phase of 0.5.

"Which is very interesting," she continued, "because it shows that gaits have symmetries in time as well as space. After all, a symmetry is just a transformation that leaves the system looking the same afterward as it was before. Periodicity itself is a time symmetry. If you shift time by one period, everything looks the same. The human walk has a mixed spatio temporal symmetry: reflect left to right and shift phase by 0.5. Isn't that grand?"

"When you were hopping, the relative phase was zero?" asked Tarzan tentatively.

"Right. You've got it."

The ape-man jumped to his feet, performed a curious dance and crashed to the ground. "I was trying to get a relative phase of 0.3," he explained." All I need to do is make my left foot lag behind my right by 0.3 of a period."

"I'm not sure you can," Jane asserted, "because it isn't a true symmetry. You see, if everything looks the same after swapping left and right and shifting phase by 0.3, then not only must your left leg be 0.3 out of phase with your right, but your right must also be 0.3 out of phase with your left. So the right leg is $0.3 + 0.3$—or 0.6 out of phase with itself, which is kind of silly."

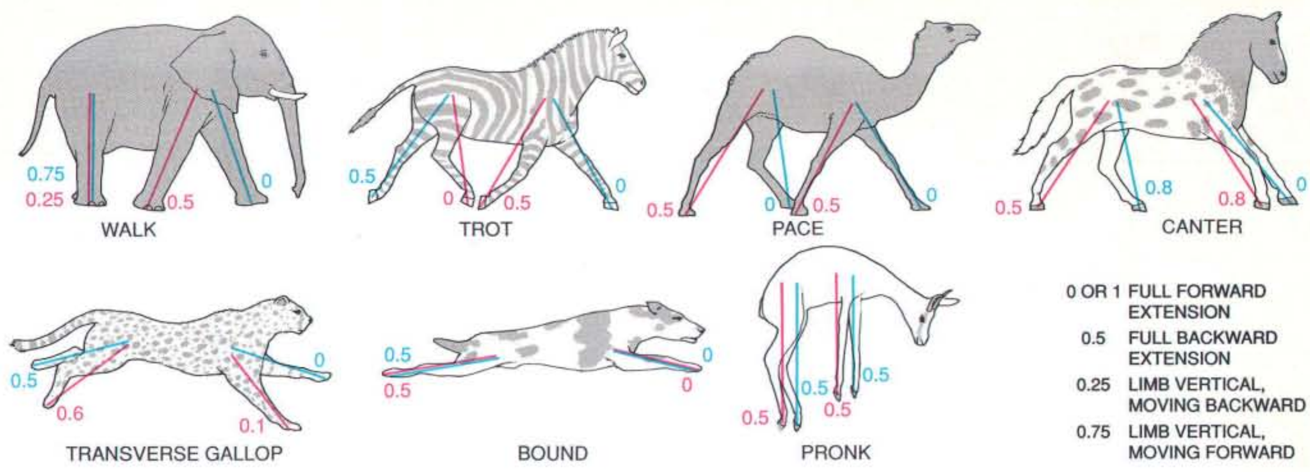
"Dangerous, too," Tarzan added, ruefully rubbing his sore leg.

"Hey! There's a theorem in all this!" exclaimed Jane. "If left-right reflection combines with a phase shift to yield a symmetry, the phase shift must be either 0 or 0.5. Nothing else is possible."

"Why?"

"Because the same argument applies. If each leg is delayed relative to the other by some phase, then each leg is delayed relative to itself by twice that phase. Now, it is possible for a leg to be delayed relative to itself—but only by an integer multiple of the period, because that's effectively the same as no delay at all. So twice the phase shift is 0, 1, 2, 3 and so on, which implies that the phase shift is 0, 0.5, 1, 1.5 and so on. But 1 has the same effect as 0, and 1.5 has the same effect as 0.5, because of periodicity.

"Which means," she went on, "that the gaits of two-legged animals can have only those two symmetries. Apart from no symmetry at all. I wonder if



Seven basic gaits, showing the relative phases of the legs

that can actually" Tarzan limped toward her, dragging one leg. "That's it, exactly! You do catch on quickly, Tarzan."

"What I don't understand, Jane, is why Curie's principle doesn't work. Why are gaits less symmetric than the whole animal? Why don't all animals hop on all fours like a pronking gazelle?"

Just then an elephant ambled through the glade, trumpeting a greeting to his human friends. "Mind you," Tarzan continued, "I don't think nature would ever allow a pronking elephant to evolve."

"Perhaps not. But Curie's principle really fails because of symmetry breaking," Jane said.

"What's symmetry breaking?"

"It happens when a symmetric system behaves in a less symmetric way. But the important point to understand is that Curie's principle can fail. Let me show you how. Where's Jim?"

Young Jim Panzee was usually hanging around near the hut—most often trying to steal bananas. Jane caught the precocious primate with ease. She then tied a knot at the end of a vine and encouraged Jim to cling to it by offering him a large, ripe banana.

"Observe that when Jim sits still and the vine hangs vertically downward," Jane pointed out proudly, "the entire system has circular symmetry." Tarzan looked baffled. "I mean, if you walk around it, it looks pretty much the same from all directions." Tarzan inspected Jim's face, then walked around to the far side. He looked more puzzled than ever. "You have to pretend Jim is a featureless lump, Tarzan." The ape-man nodded happily.

"Now suppose I grab the vine where it's draped over this branch and pull it up and down. The important part of the system, the bit of vine hanging from the branch with Jim attached, is still circularly symmetric, even though it bobs up and down. But look what happens." As Jane pumped the vine, Jim began to swing in an arc, short at first, then longer and longer. The chimp squealed in delight, waved his arms and fell off, ending the experiment.

"I saw it," Tarzan said "but I'm not sure what I saw."

"Symmetry breaking," Jane declared. "The perfectly symmetric state of the system is to hang vertically. When I pump it, the system becomes unstable. Although in theory the unstable, symmetric state can persist, you don't observe it in practice because any tiny random deviation tends to grow. Since the perfectly symmetric state can't occur, then naturally the system has to do something else, which perforce has to be less symmetric."

"I see."

"But, it's not totally asymmetric. Jim was swinging to and fro in a plane. If you think of that plane as a mirror, then his swing is symmetric under reflection. That's an example of a standing wave. And that's not all." She picked up the chimp, calmed him down with another banana and attached him to the vine again. "There's another type of oscillation that Jim can perform, too." She gave the ape a shove, and he swung in circles. "You might think this motion has circular symmetry, but that's not true. If you rotate the system through some angle, then it doesn't look exactly the same."

"No, it's like the walk in a mirror. It's the same general kind of motion, but in a different place at a given time."

"Right. What does that mean?"

"The timing's wrong—of course. It's a phase shift again."

"You've got it, Tarzan. If you rotate the system, and apply a suitable time delay, it looks exactly the same as before. And in this case, the time delay is the same as the rotation, in the sense that a rotation of 0.4 of a turn needs a time delay of 0.4 of a period, and so on. It's called a rotating wave."

"When the perfectly symmetric state becomes unstable," Tarzan reasoned, "The symmetry can break to either a standing or a rotating wave. The standing wave has a purely spatial symmetry—reflection in its plane. The rotating wave has a mixed spatio temporal symmetry."

"That's it, exactly!" Tarzan beat his chest and howled triumphantly, while Jane shook her head in disapproval. The ape-man's education still had some way to go.

"But, the circular symmetry hasn't totally vanished," Jane asserted, grabbing the vine quickly and startling Jim. "Choose a vertical plane passing through the top of the vine."

"In line with that tree over there," Tarzan answered. Jane pushed Jim towards it; the chimp oscillated to and fro in the plane that Tarzan had chosen. "Which planes will that work for?"

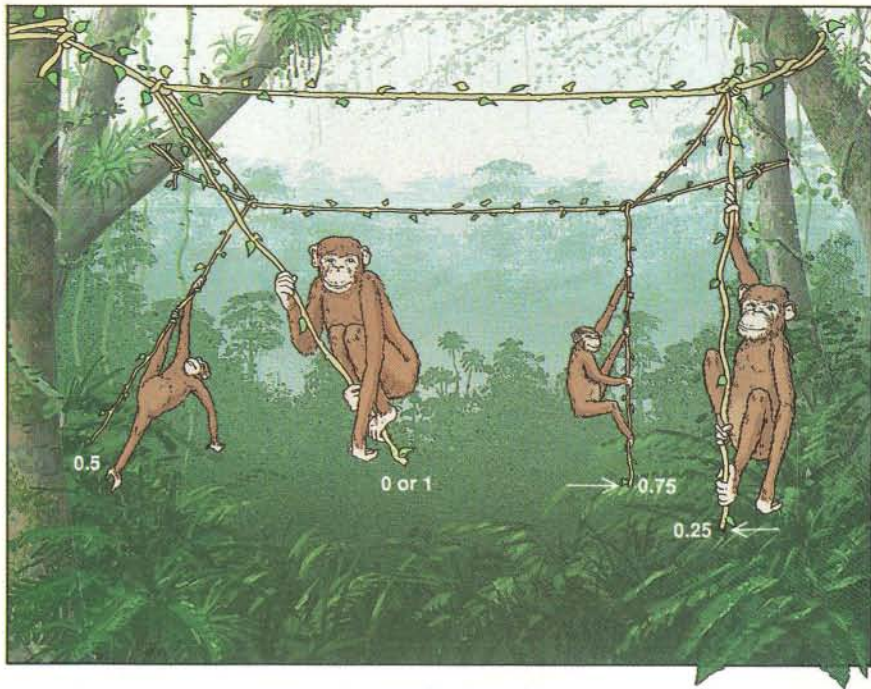
"Any of them, I guess," replied the ape-man. "Provided they're vertical and run through the point where the vine runs over the branch."

"Right. They're planes through the symmetry axis. And how are those planes related?"

"Hmmm. They're all rotations of one another. I see! Instead of having a single state of the system, unchanged by all rotations—that is, a fully symmetric state—you get lots of less symmetric states, all related to one another by rotations."

"Exactly. The whole set of motions still has circular symmetry, in the sense that if you rotate any motion, you get another one in that set. The symmetry isn't so much broken as shared."

At that moment a spotted yellow



Tarzan's central pattern generator

beast shot across the clearing, roaring. It jumped on Tarzan. There was a brief scuffle, from which the ape-man emerged wearing a broad smile and cradling a large cheetah. "Look, Spot's come to visit!"

"Yes, and using what I judge to have been a transverse gallop," observed Jane, "which is one of the least symmetric gaits."

"What symmetry does it have?" asked Tarzan.

"You can read it off from the phase shifts," Jane responded. "In the transverse gallop, diagonally opposite legs are 0.5 out of phase. There's also a curious phase lag of 0.1 between the front left and front right leg, which I'm not going to explain, except to say that it's probably related to the efficient use of energy by the animal. The symmetry is this: interchange diagonal pairs of legs and shift phase by half a period" [*see illustration on preceding page*].

"What kind of symmetry breaking could create that kind of motion?" Tarzan wondered. But the setting sun was setting. They retired to their hut.

The next morning Jane was roused by raucous screeching and chattering. Looking into the clearing, she saw Tarzan and a pack of monkeys. Tarzan had rigged up a complicated network of vines between four trees [*see illustration above*] and was trying to use bananas to bribe some young chimps into clinging to the ends of four hanging vines.

"It's a model of what the biologists call a central pattern generator," Tarzan said happily. "Each chimp represents a component of the animal's neural circuitry that controls a leg. The vines are interconnections that couple the neurons together, so that they affect one another. The system's dynamics controls the rhythms of the gait. Watch!" He gave one chimp a shove, and the animal began to swing; the impulses transmitted along the linking vines soon set the other chimps swinging in sympathy. A rather complex pattern was just setting in when one chimp jumped off to steal another's banana.

"Just a hardware problem," the ape-man explained. "Each network permits a whole range of oscillations. That's why an animal can employ several different gaits, depending on speed, terrain and so on. I can simulate most of the standard gaits using a square arrangement of vines. Oddly enough, the one that I can't seem to reproduce is the four-legged walk. That's a kind of figure-8 rotating wave, in which the front left, back right, front right and back left legs move in sequence, with 0.25 phase lags. But I can simulate that if I rearrange the vines to make two of the side connections cross."

"Let me see if I understand what you're suggesting," Jane said. "You're looking at networks of coupled oscillators and finding out what kinds of symmetry breaking can occur. Then you're trying to match the results with the actual gaits, on the assumption that each leg is controlled by one oscillator."

"Yes. Although each oscillator, as you call it, could be a complicated circuit in practice. And it works! Look, suppose you want a bounding gait. Then you set the front two 'legs' moving together and at the same time"—he rushed to the other end of the clearing—"you set the other two legs going together, but 0.5 out of phase. Of course, you can start the chimps swinging in any pattern you like, but only a few patterns persist for very long. The rest get all muddled. So I figure those are the natural oscillation patterns of the network. It's just as easy to get the trot, the pace and the pronk.

"The two types of gallop aren't so much harder, but I'm having real trouble persuading these chimps to canter, I can tell you! Probably just need more bananas to fix the problems."

"Tarzan, that's really rather impressive—"Jane began, but the ape-man had dived into the bushes, shouting, "Bugs! It ought to work for bugs too!" He reappeared waving a green beetle, and placed it on a rock. After a hesitant start, the insect scuttled off.

"Tripod gait," Jane observed. "Legs go together in threes, one triple being 0.5 out of phase with the other. Nice symmetries." By evening Tarzan had rigged up a set of six vines linked in a hexagon, and six puzzled chimps were swinging in a tripod gait. Three chimps on one side swung 0.5 out of phase with three on the other.

Just after sunrise the next day, Jane awoke to the most appalling racket she'd ever heard. Peering outside, she saw Tarzan chopping down trees to make a track through the jungle. At one end of the track lay a huge pile of vines; at the other, a heap of bananas as large as their hut. Chimpanzees were charging around everywhere.

She wondered why Tarzan would enlist the help of 100 chimps? The answer hit her like a pronking elephant: centipedes.

FURTHER READING

SYMMETRICAL GAITS OF HORSES. Milton Hildebrand in *Science*, Vol. 150, pages 701–708; November 5, 1965.

HOW MAMMALS RUN: ANATOMICAL ADAPTATIONS. P. Gambaryan. Wiley, 1974.

NEURAL CONTROL OF RHYTHMIC MOVEMENTS IN VERTEBRATES. Avis H. Cohen, S. Rossignol and S. Grillner. Wiley, 1988.

A SYNERGETIC THEORY OF QUADRUPEDAL GAITS AND GAIT TRANSITIONS. G. Schöner, W. Y. Jiang and J. A. S. Kelso in *Journal of Theoretical Biology*, Vol. 142, No. 3, pages 359–391; February 9, 1990.

BOOKS

A physicist's life examined, how to look at satellite images, a well-loved submersible

by Philip Morrison

CHANDRA: A BIOGRAPHY OF S. CHANDRASEKHAR, by Kameshwar C. Wali. University of Chicago Press, 1990 ($29.95).

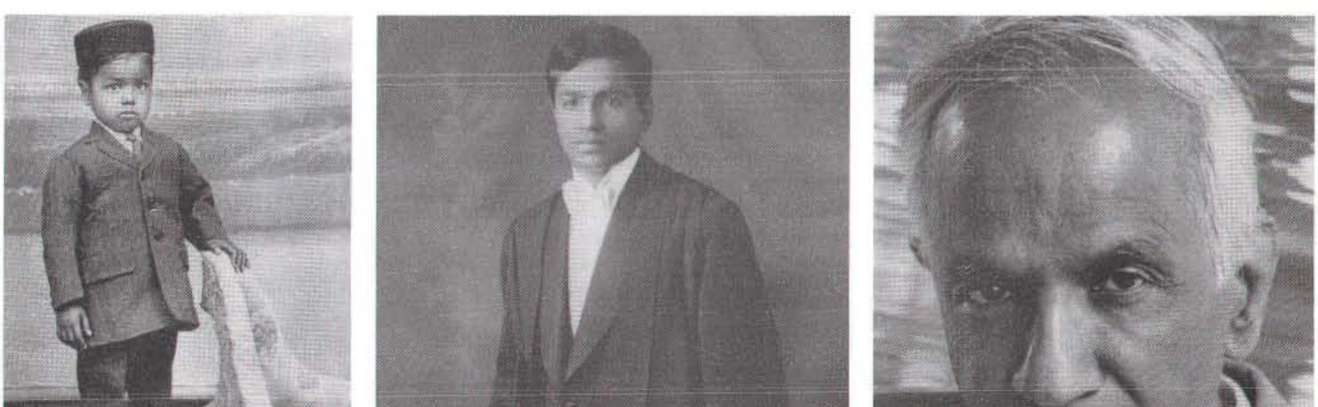
This celebrated professor is white-haired now and a little frail at 80, elegant in appearance, graceful and gentle of gesture and speech, unfailingly meticulous (as well in making pancakes as in forming equations), matchless and prolific in sustained analysis, devoted alike to teaching and research. His mother described his qualities to her friends to justify her support of his leaving home, not yet 20, for study in England. Third child among 10, her brilliant and delightful eldest son, Subrahmanyan, would be sorely missed by an ailing mother. That both grace and knowledge were so showered on him marked "only one in a million," she said. "He is born for the world, not for me. I can't come in his way.... That is the single gift that a mother can give." Mother and son were never again to meet; she died during his first year out of India, her gift to the world final.

Professor Wali, himself an American physicist of Indian origin, was a young admirer of Chandra's who became his friend. He does not try to appraise the abundant theoretical harvest but instead to portray the living man and his career. These pages muster no equations; there are period photographs and memories and taped opinions gathered from the archives and from some 50 interviews over a dozen years. We hear from colleagues and students and in particular from his wife, Lalitha, who graces another exceptionally talented and courageous Madras family. (The two met while both were teenage physics students in Madras.)

When young Chandra sailed off that July of 1930 to enter graduate study in Cambridge, he had with him on shipboard three standard texts and one new research paper of his own. That paper was a recalculation of the density of a white dwarf star, an improvement over the path-breaking estimate made by Professor Ralph H. Fowler, with whom he hoped to work. It suddenly occurred to the student that at those high densities relativity might become important. Once the sea turned smooth, he could work out the consequences. They were revolutionary: what he found was the famous "Chandrasekhar limit," now known to be written in the sky in the masses of neutron stars, each one the cooled core of some awesome supernova.

But the voyage among the astronomers was not to turn smooth quickly. Sir Arthur Eddington, never privately unfriendly to the new student, time after time humiliated him by publicly denying the reality of that limit. In vain Chandra extended and elaborated his calculations; in vain Paul Dirac and Wolfgang Pauli agreed that the limit was certainly right; in vain did Rudolf Peierls publish a new derivation aimed squarely to refute the critique. Yet no one would stand up to contradict the authority of Sir Arthur, even though the young man's theory was all but universally accepted. At the bottom was Eddington's uncritical identification with his own new "fundamental theory." Physicists had simply stopped taking him seriously, but they were not yet willing to proclaim their flat dismissal of the once great man. To this day, Chandra feels the injury to himself and the damage to science. The hurt came not from any animus on Eddington's part but from his self-assured and "aristocratic view of science," by this time blurred into dream.

Chandra's was a remarkable family at a remarkable time and place. His father, a man of originality and firmness of character, was a civil servant of high rank and reputation; his mother lovingly served her household yet had found time to translate Ibsen's *A Doll's House* into Tamil. His father's somewhat younger brother, C. V. Raman, was an extraordinary experimental physicist who in 1930 was the first Indian scientist to earn the Nobel Prize. With the prize, Raman's keen sense of his gen-



CHANDRA at the ages of six, 23 and 73.

uine importance to Indian physics was heightened, culminating eventually in his establishing the Raman Institute. Soon after winning the prize, Uncle Raman offered his stellar nephew a position at the Indian Institute of Science in Bangalore. Chandra's father received a copy of that letter of invitation. His cable to an astronomical son was crystalline: "MY ADVICE KEEP OFF HIS ORBIT."

Chandra went off instead to the New World, in 1936 settling at Yerkes Observatory of the University of Chicago. In Chicago he has remained, a glow in our astronomical sky, a leader of theory in a comprehensive style not approached since Lord Rayleigh's day. Style has always been important to this scientist. He holds the beauty of logical form and language higher than do most physicists, opportunists all, "I practice style...not only just reading... T. S. Eliot, Virginia Woolf, and Henry James, but also by paying attention to how...they construct sentences and... paragraphs." His is "a quest after perspectives"; once he has gained a view of his own, "I have the urge to present... a coherent account with order, form, and structure." That seamlessness is not without its price, even given the heroic concentration that so distinguishes the man. The world's fabric is close-knit, yet not every strand can be woven by the wisest of theorists; deeper perspectives still may emerge from unexpected novelty.

The bookshelves of most working astrophysicists carry one or another of Chandra's half-dozen big monographs, his first on stellar structure in 1939, his latest in the 1980s on the mathematics of black holes. Each monograph sketches an architecture subfield by subfield. The long series of his linked research papers serves as well-carved masonry for those constructions.

Chandra's devotion to public service is legendary, from his eager participation as a new citizen in our national elections to the engaged editorship of our chief astronomical journal, where for a generation he touched every paper for the better. The tale is true that he drove back and forth 100 wintry miles each class day between the isolated observatory and the city campus to make sure he did not miss the distant lectures for which only two graduate students had formally registered. (It does not spoil the story either to learn that those students were Chen-Ning Yang and Tsun-Dao Lee, two future Nobel laureates.)

This modest and moving biography closes with an intimate, rather triste colloquy in which Chandra, in open discontent with a science become so hasty and competitive, reflects that he still worries whether a paper he sends off will be accepted. For all his authority and renown, contentment has somehow escaped the scholar who for so long and so well sought the simple and the true, perhaps ultimately beyond reach. One wise physicist appraised Chandrasekhar in a way that helps us to understand and even to admire that touching discontent: "There is a kind of fineness about him," both physically and philosophically, as there has been ever since childhood, then for a loving family and among admiring friends in Lahore and Madras.

A GUIDE TO REMOTE SENSING: INTERPRETING IMAGES OF THE EARTH, by Stephen A. Drury. Oxford University Press, 1990 ($85; paperbound, $39.95). **THE EARTH—FROM SPACE: A SATELLITE VIEW OF THE WORLD**, by Tom Van Sant and the Geosphere Project of Santa Monica, Calif. Poster sheet in color, 24 by 36 inches. Spaceshots, Inc., 1990 (P.O. Box 3430, Manhattan Beach, CA 90266; telephone orders: 1-800-272-2779) (paper, $15; laminated, $20).

Orbital images are commonplace marvels today. Many atlaslike volumes array fascinating images but focus on what is in the picture. This enticing book opens to the general reader a wide window on understanding the imaging process. Its author is a geologist at the Open University in Milton Keynes, England, an experienced interpreter of images and a first-rate teacher, who seeks "to draw readers into *using* remote sensing" as every day it grows more relevant to global needs. We learn from this copious set of "aesthetically compelling and intellectually stimulating pictures" not only what is to be seen but even more how to see it.

Examples? One grayish image is dappled by small white spots. This is imaging radar, shuttle-borne over the plains of China's Yellow River. Each spot is a village, almost one for each square kilometer. In Minnesota it is not hard to find similar broad wheatlands but with only a single farmhouse, not one village, to each square mile. The life of a billion Chinese on the farm begins to show from orbit. Why do those villages and their irrigation canals shine so brightly under radar illumination from the side? They abound in hard, angular surfaces; every mud wall is a fine reflector. A set of well-made diagrams explains the geometry beautifully.

The usual false reds of vegetation are nicely explained; chlorophyll reflection peaks in the near infrared to keep the leaf's photosynthetic structures cool under the hot sun. Of course the leaf looks bright to a flying video camera sensitive in the IR. Code the IR in red, and you lose no other color information. The scheme was war-born, from the aerial photography of World War II; nearly the entire technology used in orbit has grown from military origins. The military worldwide are still its main users; the smaller states are the eager purchasers of publicly available civilian products from the spacefaring powers.

Look at the Shatt-al-Arab in 1977, carrying to the sea the merged waters of old Tigris and Euphrates. Here is the same scene in 1986; large artificial lakes and channels appear that served to delay the Iranian infantry attack aimed at Basra city. Iraqi hydrological engineering on large scale and its coverage from orbit are both likely to be even more notable during the 1990s.

A test image shows a grating of lines, their contrast increasing toward the top of the picture, and the line space decreases to the right. Your eye can pick out a particular grating space with lower contrast than the eye requires for either wider or narrower line spacing. The computer can take the hint: replace each eight pixels by one that bears the average value of their brightness. Fine detail will fade away. Subtract that result from the original numerical image; the new image will be all hard edges. Add it back into the original. A step-by-step example is given.

Numbers are versatile; perform by extension such tricks with any digital image, especially a set taken in multiple wavebands, and a complex subdivision of any scene can be made. Find out "ground truth" at some particular place, the real vegetation or rock type or whatever it may be, ask the computer to link any near-matching run of the pictorial digits in a single color, and you have the basis for detailed classification maps of any area. Farm crops are one example shown. It took only a week to map 1,000 square kilometers field by field. Even tonnage at a harvest yet to come can be estimated, given good records.

There are ambitious plans for an International Space Station. They call for three or four large, long-lasting satellites, crossing the equator morning and noon, in near-polar orbits, able to dip low every few years for service calls by astronauts. They would be crammed with systems to image down to 100-foot resolution, measure the height of sea and ice to an inch or so, find winds and snow and soil moisture, keep track

of chemical constituents of the air, even use laser ranging to track tectonic plate motion. What maps we will one day see!

The piper must be paid. It is true that military intelligence spends plenty of money now on those real-time Keyholes with their six-inch resolution. The reassurances of strategic change and the wars on a smaller scale that now frighten us may or may not give way to a millennial dividend of data as the cold war fades. Privatized, our costly Landsat no longer offers the concessionary prices that were once American policy. SPOT-IMAGE of Toulouse is even more "businesslike," getting nine times Landsat's price per scene. Nor can you publish or exchange SPOT data without paying serious royalties—"one reason why there are so few in this book." Military users and well-heeled multinationals after oil, minerals and news now form the ready market. The foresight of the government of India with its own Insat has put it into a better position, and that strategy may attract others. Workstations are everywhere and getting cheaper. Hope is in orbit; the International Space Station, "if it flies," offers "breathtaking possibilities."

Tom Van Sant, an artist of earth and sky, has just published with a computer-wise partner, Lloyd Van Warren, their remotely viewed physical map of the world, an assemblage of images made by TIROS multiband cameras 520 miles up. By painstaking eye-on-screen selection among alternative pixels on file, they compiled the first perfected photograph of our whole planet, everywhere on earth a cloudless summer day, although snowcaps glisten white on alpine heights. Offshore from the great deltas silt plumes stain the seas. Fields and forests are green, deserts brown, oceans blue. Here our world poses, quite still, yet sharply alive, rendered at four-kilometer resolution in just 32 million pixels.

WATER BABY: THE STORY OF *ALVIN*, by Victoria A. Kaharl. Oxford University Press, 1990 ($21.95).

This water baby is the Deep Sea Research Vessel *Alvin* out of Woods Hole, an endearing knobbly fiberglass tub weighing 18 tons in air, on its top a red conning tower like an oversized Coke tin. That blunt bow is fitted with sonar, big sample basket, and a bright light. The sub was launched in 1964—though by now little but design continuity remains from that old *Alvin*—and has for more than 2,000 times taken pilot and two passengers—scientists, trainees, even television performers—a couple of miles down for a long workday and up again. Kaharl, a science writer in residence at the Woods Hole Oceanographic Institution, has pored over the logs and the reports and the piles of recorded tapes, talked to and worked alongside the sub's devoted people of every disposition and talent, dived herself, a little frightened, and come up with an intimate chronicle of 25 years.

At the center lies perfection. It is the seven-foot passenger sphere in which the crew can live at surface pressure. That sphere symmetrically bears its fierce load; one test sphere did not implode during a whole week at 20,000 feet equivalent. (Once, under stress, the big testing tank exploded instead, throwing its 20-ton lid high in the air while the engineers ran. The lid in a matching Platonic symmetry fell back into place, shoving the whole tank three feet down into the ground.) That first sphere was of alloy steel; a photograph shows the proud welders at work in Houston, fusing its two hemispheres. "She sure looks purdy," the supervisor wrote on the photograph. The current sub has a two-inch welded titanium sphere made in the 1970s, stronger than steel and much lighter.

Ours is not Plato's world, but conspicuously imperfect on land or sea. The real sphere has big holes for the hatch and the four thick acrylic plastic portholes. It nests into that tubby outer hull where dwell the ballast pump, the steel weights to be released magnetically for sure lift, the big lead-acid batteries, and built-in buoyancy in smaller empty titanium spheres and in plastic foam. Power has to enter, and control signals must leave the passenger space; what spherical purity remains vanishes with the penetrators, a dozen little tapered sealing tubes of titanium that lead all the wires through holes in the strong shell. "Some pilots... believe that if anything is going to kill them, it will be a failed penetrator"; the steel-hard high-pressure jet of seawater that would enter could easily decapitate.

But design alone does not rule this world of action. Crew chief George Broderson, virtuoso mechanic, unfailingly theatrical and a man long head over heels in love with his submersible, tightened the new penetrator nuts himself. "What felt right to him was getting those suckers just as tight as he could...whacking the wrench with a sledgehammer." Asked by the brass for the torque setting he used, he "gravely intoned *125 foot-pounds, sir.*" That re-

mains the standard because it works, although the more anxious members of the corps of pilots still steadily taste for saltiness the condensate that drips blandly from the cold inner ends of the penetrators.

A submersible for the oceanographers was publicly proposed by Allyn Vine of WHOI in 1956; his steady advocacy found a listening ear in the Office of Naval Research. *Alvin* was named for him, and the first of its mother ships, the wonderful if snail-slow improvised catamaran that cradled *Alvin* afloat for more than a decade, was named for Vine's mother, Lulu. At sea most of the crew lived in Lulu's nearly submerged starboard steel pontoon; they called it the Tube of Doom. In just such an insider's vein the detailed narrative continues, a wide-ranging and good-humored story of a couple of thousand leagues deep under the sea at half-knot speed. The struggles of research funding and priority are here, no less than how three passengers, male and female, can tactfully manage to urinate inside that seven-foot sphere chock-full of electronics. (The first female pilot is as well the first among the 20-odd certified pilots to be a trained scientific oceanographer.)

The exploits of *Alvin* are both well sung and well explained. The sub gained credibility first when it found the H-bomb that SAC (Strategic Air Command) lost into the sea just off the Spanish coast. With its French kin, *Alvin* prowled the Mid-Atlantic Ridge, field geology done on the ocean floor for the first time. One windy morning a cable broke as the sub was being uncradled only a day's sail out of Woods Hole, and *Alvin* sunk, wave-swamped through the open hatch; the crew escaped—barely. They found and fished her up at last, again with help from a big aluminum friend; the bologna sandwich that came up fresh after 10 months is famous. Bacteria work slowly down there in the cold; the deep sea is no place to dispose of garbage.

Diving to the *Titanic* is the most popular of *Alvin* stories, as her marvelous discovery of hot-water seafloor vents and their vigorous life is the most important. The knowing author avoids the familiar overstatement of the independence of that submarine life from the sun; the molecular oxygen down there, plainly essential to all those red-blooded clams, is photosynthetic in origin. Only here and there along the rifts do some microorganisms thrive free of all oxygenic or carbonaceous debt to solar energy. Perhaps in the very beginning, microlife itself began at some ancient vent.

SCIENTIFIC AMERICAN CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at $1.25 each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, Utah 84104. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $1.50 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, Mass. 01970. Specify CCC Reference Number ISSN 0036-8733/**91**. $1.50 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017. Manuscripts are submitted at the authors' risk and will not be returned unless accompanied by postage.

Advertising correspondence should be addressed to Advertising Manager, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, N.Y. 10017.

Address subscription correspondence to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 3187, Harlan, IA. **51537**. Telephone inquiries: 1-800-333-1199, U.S. only; other 515-247-7631/32. The date of the last issue on subscriptions appears on each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, a mailing label of a recent issue) as well as the new one.

Name

New Address	Old Address
Street	Street
City	City
State and ZIP	State and ZIP

ESSAY

America's green strategy

by Michael E. Porter

Do strict environmental standards make American industry less competitive in international markets? Many observers answer yes. Richard Darman, director of the Office of Management and Budget, has quipped that "Americans did not fight and win the wars of the 20th century to make the world safe for green vegetables."

The conflict between environmental protection and economic competitiveness is a false dichotomy. It stems from a narrow view of the sources of prosperity and a static view of competition.

Strict environmental regulations do not inevitably hinder competitive advantage against foreign rivals; indeed, they often enhance it. Tough standards trigger innovation and upgrading. In my book *The Competitive Advantage of Nations,* I found that the nations with the most rigorous requirements often lead in exports of affected products.

Although the U.S. once clearly led in setting standards, that position has been slipping away. Until the passage of the Clean Air Act in 1990, itself the result of 12 years of foot-dragging, Congress had passed little environmental legislation since the mid-1970s. Today the U.S. remains the only industrialized country without a policy on carbon dioxide, and our leadership in setting environmental standards has been lost in many areas. Even Japan, a nation many think of as relatively unconcerned about the environment, has moved ahead of the U.S. in important fields. Japan's NO_x emission standards for vehicles are significantly more stringent than those in the U.S. and Europe: its stationary SO_x and NO_x standards are set in terms of rigorous daily (versus yearly) average hourly emissions.

As other nations have pushed ahead, U.S. trade has suffered. Germany has had perhaps the world's tightest regulations in stationary air-pollution control, and German companies appear to hold a wide lead in patenting—and exporting—air-pollution and other environmental technologies. As much as 70 percent of the air pollution-control equipment sold in the U.S. today is produced by foreign companies. Britain is another case in point. As its environmental standards have lagged, Britain's ratio of exports to imports in environmental technology has fallen from 8 : 1 to 1 : 1 over the past decade.

In contrast, the U.S. leads in those areas in which its regulations have been the strictest, such as pesticides and the remediation of environmental damage. Such leads should be treasured and extended. Environmental protection is a universal need, an area of growing expenditure in all the major national economies ($50 billion a year in Europe alone) and a major export industry. Without competitive technology, America will not only forsake a growth industry, but more and more of our own environmental spending will go to imports.

Even in the broader economy, strict environmental codes may actually foster competitiveness. Exacting standards seem at first blush to raise costs and make firms less competitive, particularly if competitors are from nations with fewer regulations. This may be true if everything stays the same except that expensive pollution-control equipment is added.

But everything will not stay the same. Properly constructed regulatory standards, which aim at outcomes and not methods, will encourage companies to re-engineer their technology. The result in many cases is a process that not only pollutes less but lowers costs or improves quality. Processes will be modified to decrease use of scarce or toxic resources and to recycle wasted by-products. The 3M Company, for example, estimates that its "Pollution Prevention Pays" program has saved $482 million since 1975, while eliminating more than 500,000 tons of waste and pollutants, and has saved another $650 million by conserving energy.

Strict product regulations can also prod companies into innovating to produce less polluting or more resource-efficient products that will be highly valued internationally. As a result of the U.S. proposed phaseout of chlorofluorocarbons (CFCs), for example, Du Pont and other American firms are pioneers in finding substitutes.

This is not to say that all companies will be happy about tough regulations: increased short-term costs and the need to redesign products and processes are unsettling at the least. The aversion to tough standards will be particularly strong in industries that feel threatened by international competition, as is too often the case in America today. The auto industry, for example, has been fighting mandates to improve fuel efficiency, even though meeting them could stimulate innovations that made products more competitive.

The strongest proof that environmental protection does not hamper competitiveness is the economic performance of nations with the strictest laws. Both Germany and Japan have tough regulations, and both countries continue to surpass the U.S. in GNP growth rates and rates of productivity growth. Japan has become a world leader in developing pollution-control equipment and cleaner, more efficient processes. It is noteworthy that in America many of the sectors subject to the greatest environmental costs have improved their international trade performance, among them chemicals, plastics, synthetics, fabrics and paints.

Turning environmental concern into competitive advantage demands that we establish the right kind of regulations. They must stress pollution prevention rather than merely abatement or cleanup. They must not constrain the technology used to achieve them, or else innovation will be stifled. And standards must be sensitive to the costs involved and use market incentives to contain them.

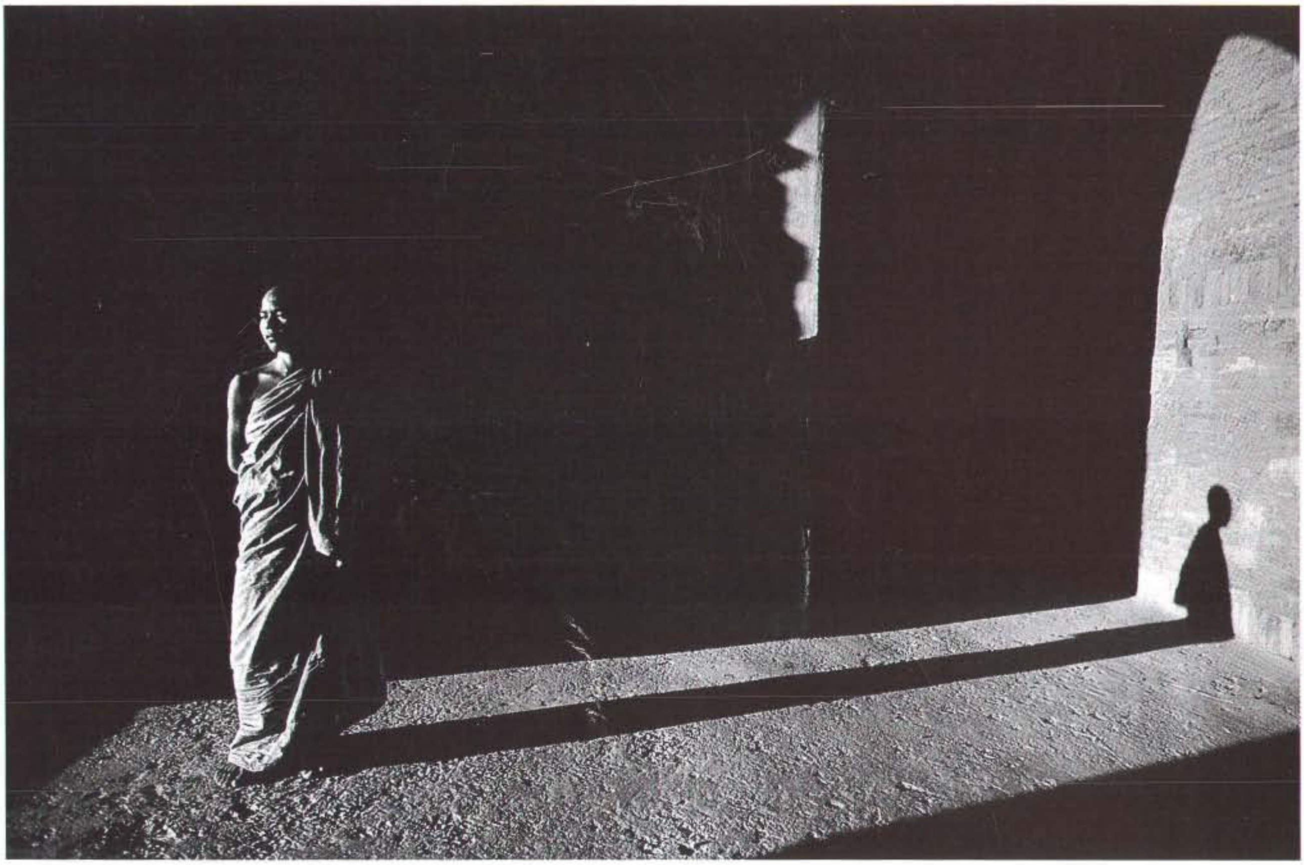
Because U.S. environmental regulations have traditionally violated these principles, the substantial amount we spend on protecting the environment has not yielded the benefits it could have. In the 1970s, for example, ambient air-quality standards encouraged tall smokestacks, some as high as 800 feet, which exported pollution somewhere else instead of reducing it. Even today most standards are met with end-of-pipe technology, where equipment is simply added to the end of a process.

The resurgence of concern for the environment, then, should be viewed not with alarm but as an important step in regaining America's preeminence in environmental technology. The Environmental Protection Agency must see its mandate as stimulating investment and innovation, not just setting limits.

In companies, the "Chicken Little" mind-set that regulation inevitably leads to costs and an adversarial posture toward regulators must be discarded. Environmental protection can benefit America's competitiveness if we simply approach it properly.

MICHAEL E. PORTER is a professor at the Harvard Business School and leads the school's Competition and Strategy group.

CONTEMPLATION



LEICA CAMERA GMBH, OSKAR-BARNACK-STRASSE 11, D-6336 SOLMS

MADE BY LEICA

The thrill of photography often lies in seeing without being seen. That means acting quickly, discreetly, and sensitively. For this, the LEICA M6 is just right. The compactness, the barely audible shutter release, the exemplary fine mechanics, and a lens range second to none, crowned by the 50 mm f/1 Noctilux, lend this camera its excellent quality and unique character. And give the photographer access to a new, fascinating dimension of photography. Leica photography.

Leica

The freedom to see.